DOCUMENT RESUME

ED 318 058 CS 507 129

AUTHOR	Pisoni, David B.; And Others
TITLE	Research on Speech Perception. Progress Report No. 14.
INSTITUTION	Indiana Univ., Bloomington. Dept. of Psychology.
SPONS AGENCY	National Institutes of Health (DHHS), Bethesda, Md.; National Science Foundation, Washington, D.C.
PUB DATE	88
CONTRACT	AF-F-33615-86-C-0549
GRANT	IRI-86-17847; NS-07134-10; NS-12179-12
NOTE	344p.; For other reports in this series, see CS 507 123-128.
PUB TYPE	Reports - Research/Technical (143) -- Collected Works - General (020) -- Information Analyses (070)
EDRS PRICE	MF01/PC14 Plus Postage.
DESCRIPTORS	*Acoustic Phonetics; Auditory Discrimination; *Auditory Perception; Communication Research; Computer Software Development; Infants; *Language Processing; Language Research; Linguistics; Speech; Speech Synthesizers
IDENTIFIERS	Indiana University Bloomington; *Speech Perception; Speech Research; Theory Development

ABSTRACT

Summarizing research activities in 1988, this is the fourteenth annual report of research on speech perception, analysis, synthesis, and recognition conducted in the Speech Research Laboratory of the Department of Psychology at Indiana University. The report includes extended manuscripts, short reports, and progress reports. The report contains the following 15 articles: "Retroactive Influence of Syllable Neighborhoods" (J. Charles-Luce and others); "Priming Lexical Neighbors of Spoken Words: Effects of Competition and Inhibition" (S. D. Goldinger and others); "Some Effects of Time-Varying Context on the Perception of Speech and Nonspeech Sounds" (J. W. Mullenix and others); "Intonational Context and FO Normalization" (K. A. Johnson); "Word Familiarity and Frequency in Visual and Auditory Word Recognition" (C. M. Connine and J. W. Mullenix); "Similarity Neighborhoods of Spoken Two Syllable Words: Retroactive Effects on Multiple Activation" (M. S. Cluff and P. A. Luce); "Similarity Neighborhoods of Spoken Words" (P. A. Luce and others); "Manner of Articulation and Feature Geometry: A Phonological Perspective" (S. Davis); "Training Japanese Listeners to Identify /r/ and /l/: A First Report" (J. S. Logan and others); "FO Normalization and Adjusting to Talker" (K. Johnson); "On the External Evidence for Y-Insertion in American English" (S. Davis); "Vowel Length and Closure Duration in Word-Medial VC Sequences" (S. Davis and W. V. Summers); "Detailing the Nature of Talker Normalization in Speech Perception" (J. W. Mullenix and D. B. Pisoni); "Manner of Articulation and Feature Geometry: A Phonetic Perspective" (K. Johnson); and "Determining the Locus of Talker Variability Effects on the Recall of Spoken Word Lists: Evidence from a Presentation Rate Manipulation" (S. D. Goldinger and others). Lists of publications and of laboratory staff and personnel conclude the report. (SR)

ED318058

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 14
(1988)

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana
47405

Supported by

Department of Health and Human Services
U.S. Public Health Service

National Institutes of Health
Research Grant No. NS-12179-12

National Institutes of Health
Training Grant No. NS-07134-10

National Science Foundation
Research Grant No. IRI-86-17847

and

U.S. Air Force
Armstrong Aerospace Medical Research Laboratory
Contract No. AF-F-33615-86-C-0549

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D.B. PISONI

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CS507129

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☐ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

BEST COPY AVAILAB...

RESEARCH ON SPEECH PERCEPTION

Progress Report No. 14
(1988)

David B. Pisoni, Ph.D.
Principal Investigator

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405

Research Supported by:

Department of Health and Human Services
U. S. Public Health Service

National Institutes of Health
Research Grant No. NS-12179-12

National Institutes of Health
Training Grant No. NS-07134-10

National Science Foundation
Research Grant No. IRI-86-17847

and

U. S. Air Force
Armstrong Aerospace Medical Research Laboratory
Contract No. AF-F-33615-86-C-0549

RESEARCH ON SPEECH PERCEPTION Progress Report No. 14
(1988)

Table of Contents

Introduction iii

I. Extended Manuscripts 1

Retroactive influence of syllable neighborhoods;
Jan Charles-Luce, Paul A. Luce, and Michael S. Cluff 3

Priming lexical neighbors of spoken words: Effects of
competition and inhibition;
Stephen D. Goldinger, Paul A. Luce, and David B. Pisoni 17

Some effects of time-varying context on the perception of
speech and nonspeech sounds;
John W. Mullennix, David B. Pisoni, and Stephen D. Goldinger . . . 47

Intonational context and FO normalization;
Keith A. Johnson 81

Word familiarity and frequency in visual and
auditory word recognition;
Cynthia M. Connine and John W. Mullennix 109

Similarity neighborhoods of spoken two syllable words:
Retroactive effects on multiple activation;
Michael S. Cluff and Paul A. Luce 125

Similarity neighborhoods of spoken words;
Paul A. Luce, David B. Pisoni, and Stephen D. Goldinger 151

Manner of articulation and feature geometry:
A phonological perspective; Stuart Davis 179

Training Japanese listeners to identify
/r/ and /l/: A first report;
John S. Logan, Scott E. Lively, and David B. Pisoni 209

II. Short Reports and Work-in-Progress 235

FO normalization and adjusting to talker;
Keith Johnson 237

On the external evidence for y-insertion
in American English; Stuart Davis 259

Vowel length and closure duration in word-medial
VC sequences; Stuart Davis and W. Van Summers 269

Detailing the nature of talker normalization
in speech perception;
John W. Mullennix and David B. Pisoni 289

Manner of articulation and feature geometry: A
phonetic perspective; Keith Johnson 307

Determining the locus of talker variability effects
on the recall of spoken word lists: Evidence
from a presentation rate manipulation;
Stephen D. Goldinger, John S. Logan, and David B. Pisoni; 323

III. Publications 343

IV. SRL Laboratory Staff and Personnel 347

INTRODUCTION

This is the fourteenth annual report summarizing the research activities on speech perception, analysis, synthesis, and recognition carried out in the Speech Research Laboratory, Department of Psychology, Indiana University in Bloomington. As with previous reports, our main goal has been to summarize various research activities over the past year and make them readily available to granting agencies, sponsors and interested colleagues in the field. Some of the papers contained in this report are extended manuscripts that have been prepared for formal publication as journal articles or book chapters. Other papers are simply short reports of research presented at professional meetings during the past year or brief summaries of "on-going" research projects in the laboratory. From time to time, we also have included new information on instrumentation and software support when we think this information would be of interest or help to others. We have found the sharing of this information to be very useful in facilitating our own research.

We are distributing reports of our research activities because of the ever increasing lag in journal publications and the resulting delay in the dissemination of new information and research findings in the field of speech processing. We are, of course, very interested in following the work of other colleagues who are carrying out research on speech perception, production, analysis, synthesis, and recognition and, therefore, we would be grateful if you would send us copies of your own recent reprints, preprints and progress reports as they become available so that we can keep up with your latest findings. Please address all correspondence to:

Professor David B. Pisoni
Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, Indiana 47405
USA
(812) 855-1155, 855-1768
E-mail (BITNET) "PISONI@IUBACS"

Copies of this report are being sent primarily to libraries and specific research institutions rather than individual scientists. Because of the rising costs of publication and printing, it is not possible to provide multiple copies of this report to people at the same institution or issue copies to individuals. We are eager to enter into exchange agreements with other institutions for their reports and publications. Please write to the above address.

The information contained in the report is freely available to the public and is not restricted in any way. The views expressed in these research reports are those of the individual authors and do not reflect the opinions of the granting agencies or sponsors of the specific research.

I. EXTENDED MANUSCRIPTS

Retroactive Influence of Syllable Neighborhoods [1]

Jan Charles-Luce,[2] Paul A. Luce,[3] and Michael S. Cluff

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This is a draft of a paper presented at the workshop on "Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspectives" in Sperlonga, Italy, May, 1988. This research was supported by NIH Research Grant *NS-12179-12* and NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN.

[2] Jan Charles-Luce is now with the Department of Communicative Disorders and Sciences, State University of New York at Buffalo.

[3] Paul Luce is now with the Department of Psychology, State University of New York at Buffalo.

Retroactive Influence of Syllable Neighborhoods

The issue addressed by Richard Shillcock and Anne Cutler (this volume) is how words, in the absence of reliable phonetic cues to word boundaries, are accessed in continuous speech. In part, both Shillcock and Cutler showed effects of non-word-initial syllables on lexical access. Shillcock demonstrated that lexical hypotheses are initiated for non-initial syllables in monomorphemic bisyllabic words. In a cross-modal priming task, he found that subjects' lexical decision times were faster when the visual probe was semantically related, versus semantically unrelated, to the second syllable. For example, when presented trombone auditorily, subjects were faster to respond when the visual probe was rib than when it was bun. Shillcock suggests that the second syllable of trombone produced an independent lexical hypothesis that primed rib, apparently before the two syllables were integrated into a single word unit. Shillcock argues that these results are not consistent with theories advocating sequential, deterministic parsing of the speech signal into words. Non-initial (stressed) syllables may initiate independent lexical hypotheses even though initial syllables may be unresolved (e.g., trom-).

Furthermore, lexical access may not always operate in a strictly left-to-right manner, and information carried in non-initial syllables may be important in helping to resolve preceding syllables. Cutler demonstrates that detecting the word mint embedded as the first syllable of a bisyllabic nonword takes longer in a Strong-Strong sequence (e.g., mintayf [mIntef]) than in a Strong-Weak sequence (e.g., mintef [mIntəf]). Cutler (Cutler & Norris, 1988) argues that in the Strong-Strong sequence (mintayf), lexical access is initiated twice–once for each strong syllable. Information from the end of the first syllable may be initially segmented as belonging to the second syllable. As a result, mintayf is initially segmented as min-tayf. When no word is found corresponding to the input of the second syllable, the listener must "reassemble" the missegmented speech signal to access the word mint. Consequently, detection times for mint are slowed in the Strong-Strong sequence as compared to the Strong-Weak sequence, where the initial segmentation is mint-ef. According to Cutler's model, weak syllables do not initiate segmentation. Missegmentation does not occur and, hence, no competition for [-t-] between syllables occurs in mintef. Thus, when the bisyllabic word is Strong-Weak, only one lexical hypothesis is initiated because there is only one strong syllable. The mintayf vs. mintef results show that non-initial syllabic contexts may affect resolution of preceding ambiguous syllables in lexical access. Strict left-to-right models of word recognition are not sufficient when missegmentation occurs and can only be successful if they allow some sort of principled "backtracking" during hypothesis matching.

Thus, both Shillcock's and Cutler's findings, in part, suggest that lexical access may not always be successful if it is a strictly sequential and deterministic process. Moreover, there is some evidence that non-initial syllabic context may be significant in the resolution of a first syllable. However, retroactive resolution may depend not only on such factors as strong versus weak syllables, but also on the neighborhood structure of the individual syllables. To demonstrate this point, consider the results from a perceptual identification task investigating spliced and natural spondees (Cluff & Luce, 1988). A spondee is a bisyllabic, bimorphemic word composed of two monosyllabic, monomorphemic words. In English, spondees are compound words composed of varying parts of speech, for example adjective + noun (lighthouse), noun + noun (fencepost), or verb + verb (hearsay). Note that spondees, while generally receiving primary stress on the first syllable, are metrically composed of two strong syllables. The second syllable is never phonetically reduced and, therefore, is never a weak syllable. Cluff and Luce (1988) found that the identification of an initial syllable occurring in a high density, high frequency neighborhood can be facilitated by a second syllable occurring in a low density, low frequency neighborhood.

Research by Luce (1986; see also Luce, Pisoni, & Goldinger, this volume) has shown that identification of monosyllabic words is affected by stimulus word frequency, neighborhood density, and neighborhood frequency. Neighborhood density is defined as the number of words that are phonetically similar to the target word. Thus, words may be in dense neighborhoods with many phonetically similar words or in sparse neighborhoods with few phonetically similar words. Neighborhood frequency refers to the frequency of the words within the neighborhood. Words may be in a neighborhood with high frequency neighbors or in a neighborhood with low frequency neighbors. In a perceptual identification task, Luce (1986) found that performance increased as stimulus word frequency increased and decreased as neighborhood density and neighborhood frequency increased. However, these results were based only on CVC monosyllabic words. The purpose of the spondee investigation was to extend the findings from the monosyllabic words to bisyllabic words. In particular, the purpose was to determine how the neighborhood characteristics of each syllable of a spondee affect lexical access.

As already mentioned, spondees in general receive primary stress on the first syllable. However, in the first set of results discussed below, stress was controlled for by digitally splicing together monosyllabic words that correspond with the two syllables of naturally occurring spondees. For example, for the spondee lighthouse, light and house were first recorded as individual words and then digitally spliced together using a waveform editor. These stimuli will be referred to as spliced spondees. The initial intent, then, was to determine the effects of neighborhoods on syllables while controlling for possible stress differences across syllables.

To gauge the effects of neighborhood density and neighborhood frequency, four sets of spliced spondees were created by manipulating independently the neighborhoods of each monosyllable of the spondees. A monosyllable was either "hard" or "easy." "Easy" signifies that the monosyllable is a high frequency word in a low-density, low-frequency neighborhood. "Hard" signifies that the monosyllable is a low frequency word in a high-density, high-frequency neighborhood. Thus, a spondee may have one of four bisyllablic patterns: EASY-EASY, EASY-HARD, HARD-EASY, or HARD-HARD. Examples of each of the four pattern types are presented in Table I.

Insert Table 1 about here

Thirty-six spondees for each pattern type were auditorily presented with a background of white noise at a signal-to-noise ratio of +5 dB. The subjects' task was to type the spondee they thought they heard.

Figure 1 shows the percent correct identification from the spliced spondee condition. Each bar represents the mean percent correct averaged across first and second syllables for each pattern type.

Insert Figure 1 about here

Results of percent correct identification showed that subjects were best at identifying easy-easy spondees and worst at identifying hard-hard spondees. Thus, performance is best when both syllables are in low-density, low-frequency neighborhoods. Performance is worst when both syllables are in high-density, high-frequency neighborhoods. Furthermore, notice that identification is better for the hard-easy spondees than for the easy-hard spondees, suggesting differential effects of neighborhood structure on first and second syllables. In order to determine the locus of this asymmetry between hard-easy and easy-hard spondees, identification performance for the individual syllables making up the spondees was examined. These results are shown in Figure 2.

Insert Figure 2 about here

Looking first at the data for the second syllables, there was a systematic effect of neighborhood structure. Easy second syllables are identified more accurately than hard second syllables, regardless of the neighborhood structure of the first syllable. This was not the case, however, for the first syllables. Identification of easy first syllables drops when second syllables are hard compared to when second syllables are easy. On the other hand, identification for hard first syllables improves when second syllables are easy relative to when they are hard.

The results from the spliced spondees suggest that there is some retroactive influence from the second to the first syllable in identification performance. Most interesting, an easy second syllable appears to help identification of a hard first syllable. This suggests that a syllable in a low-density, low-frequency neighborhood helps resolve preceding ambiguity of a

Table I. Examples of spondee patterns.

EASY-EASY:	jigsaw
EASY-HARD:	causeway
HARD-EASY:	bucksaw
HARD-HARD:	hearsay

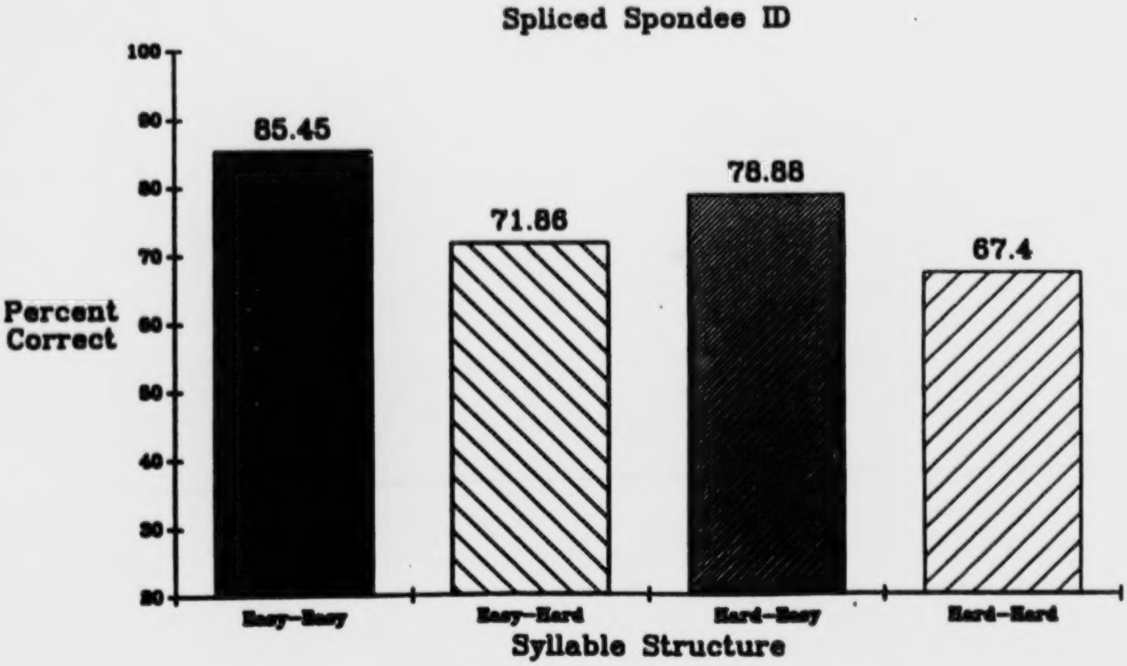


Figure 1. Percent correct identification for spliced spondees.

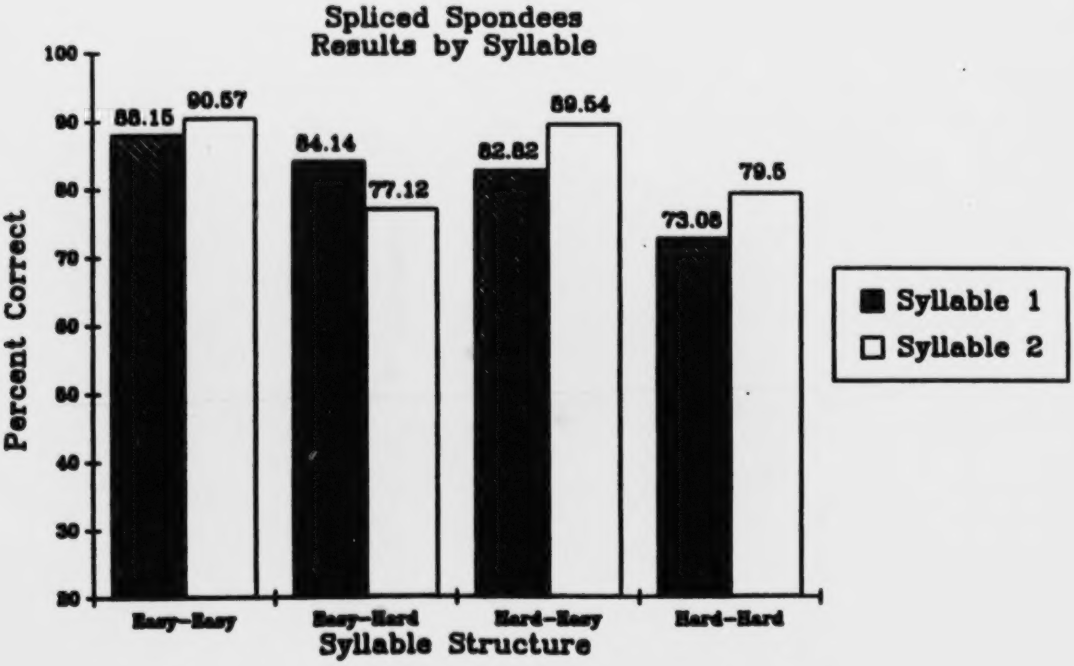


Figure 2. Percent correct identification for the individual syllables in the spliced spondee condition.

syllable in a high-density, high-frequency neighborhood.

Recall that stress was controlled in the spliced spondees–each syllable of the spondee was produced as a single word and these words then were spliced together digitally to construct the spondee. The next two figures show the results for the naturally produced spondees. The same four patterns and the same 36 spondees in each pattern type were used. Thus, naturally produced EASY-EASY, EASY-HARD, HARD-EASY, and HARD-HARD spondees were presented for identification. The task was identical to the spliced spondee condition.

The results for the naturally produced spondees are presented in Figure 3. Percent correct identification averaged across first and second syllables are given for each pattern type.

Insert Figure 3 about here

The results from the naturally produced spondees replicate the patterns of results from the spliced spondee condition. When both syllables are in low-density, low-frequency (easy) neighborhoods, identification is better than when both syllables are in high-density, high-frequency (hard) neighborhoods. Also, identification is better when the second syllable is easy compared to when it is hard.

Although the pattern of results is consistent with the spliced condition, overall identification for the natural spondees is somewhat attenuated from the spliced spondees. The overall decrement in performance is due entirely to the second syllable. Percent correct is virtually identical for the first syllables for both the spliced and natural spondee conditions. This attenuation may be due to the fact that the second syllable does not carry primary stress. One consequence of this might be that the second syllable is less stressed and, therefore, somewhat shorter in duration in the natural condition than the spliced condition, where the second syllable was produced as an isolated word. However, measurements of each syllable showed that the mean duration of the second syllable is not shorter than the first syllable, irrespective of the syllable neighborhoods. Therefore, some other acoustic parameter associated with the production of the second syllables must explain the attentuation in performance (e.g., less clearly articulated overall, greater diphthongization, lower amplitude or pitch, among others).

Nonetheless, the results from the natural spondees again show that the neighborhoods of the first syllables have little affect on the identification of the second syllables. Figure 4 shows percent correct identification for the individual syllables of each pattern type for the natural spondee condition.

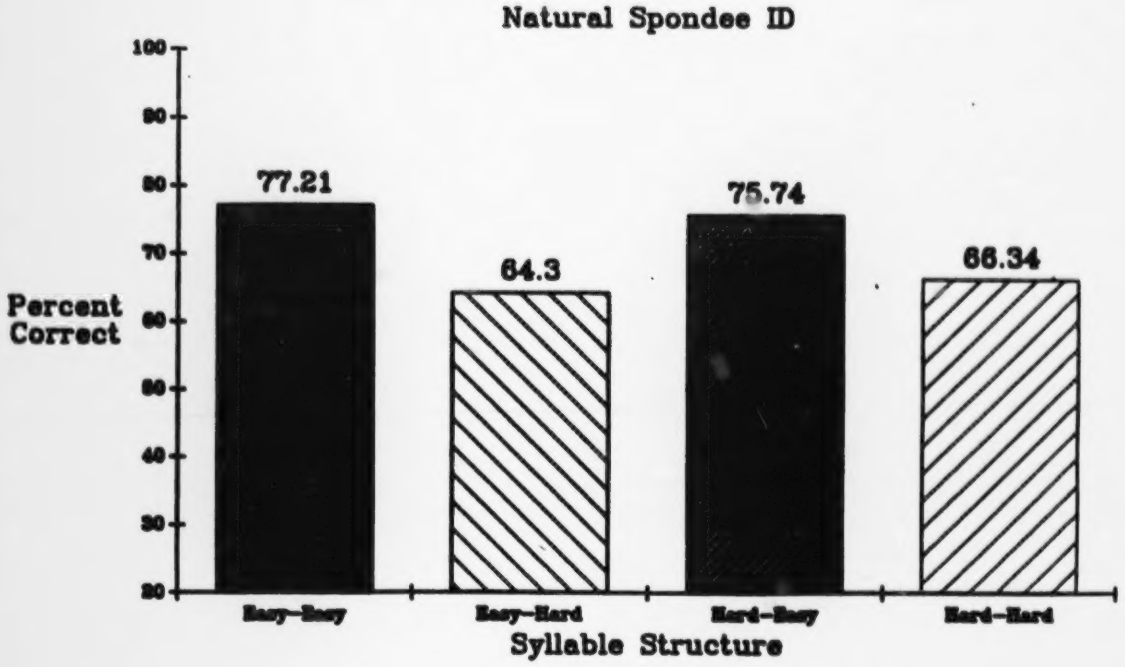


Figure 3. Percent correct identification for natural spondees.

Insert Figure 4 about here

Second syllables in low-density, low-frequency (easy) neighborhoods are identified more accurately than second syllables in high-density, high-frequency (hard) neighborhoods, regardless of the neighborhoods of the first syllables. Furthermore, and again similar to the spliced spondees, identification is better for the hard-easy spondees than for the easy-hard spondees, reflecting the asymmetrical effects of neighborhood structure on first and second syllables. In particular, note that identification increases for a hard first syllable when the second syllable is easy.

To summarize the spondee data, there appears to be differential effects of neighborhood structure on first and second syllables. An asymmetry in identification performance was found between hard-easy spondees and easy-hard spondees, regardless of the splicing. In general, spondees represent a Strong-Strong metrical syllable structure because of their primary-secondary stress pattern and not a primary-reduced stress pattern. The second syllable never completely reduces such that the vowel quality is schwa (ə). Assuming Cutler's model of lexical access (that is, strong syllables initiate lexical hypotheses), we suggest lexical hypotheses are initiated independently for each syllable of the spondee (see Cutler & Carter, 1987; also Grosjean & Gee, 1987). Thus, not only is there competition among neighbors for individual syllables of spondees, but there appears to be some retroactive effects from the easy second syllable to facilitate access of the hard first syllable from among its neighbors.

Therefore, overall success of accessing a polysyllabic word may depend upon the composition of the word in terms of neighborhood characteristics of each syllable. Obviously, however, not all syllables will be complete words in and of themselves, as are the syllables composing a spondee. Nonetheless, it is not unreasonable to hypothesize that for such bisyllabic, monomorphemic words like trombone that two lexical hypotheses (one for each of the strong syllables) are initiated. This again follows from the evidence suggesting that strong syllables bootstrap lexical access.

Furthermore, the neighborhoods activated by each strong syllable may interact in such a way as to facilitate the correction of erroneous lexical hypotheses. For example, in trombone the two independent lexical hypotheses initiated by each strong syllable, trom- and -bone, must be resolved at some point to access the single word unit trombone (see Shillcock, this volume). Depending upon the neighborhoods activated by each syllable, lexical access of a bisyllabic word may be facilitated retroactively if the second syllable is in a low-density, low-frequency neighborhood. Access may not be facilitated if the second syllable is in a high-density, high-frequency neighborhood.

With regard to Shillcock's pseudo-prefixed words, the initial syllable is phonetically reduced and therefore a weak syllable. These words are "pseudo"-prefixed because, although

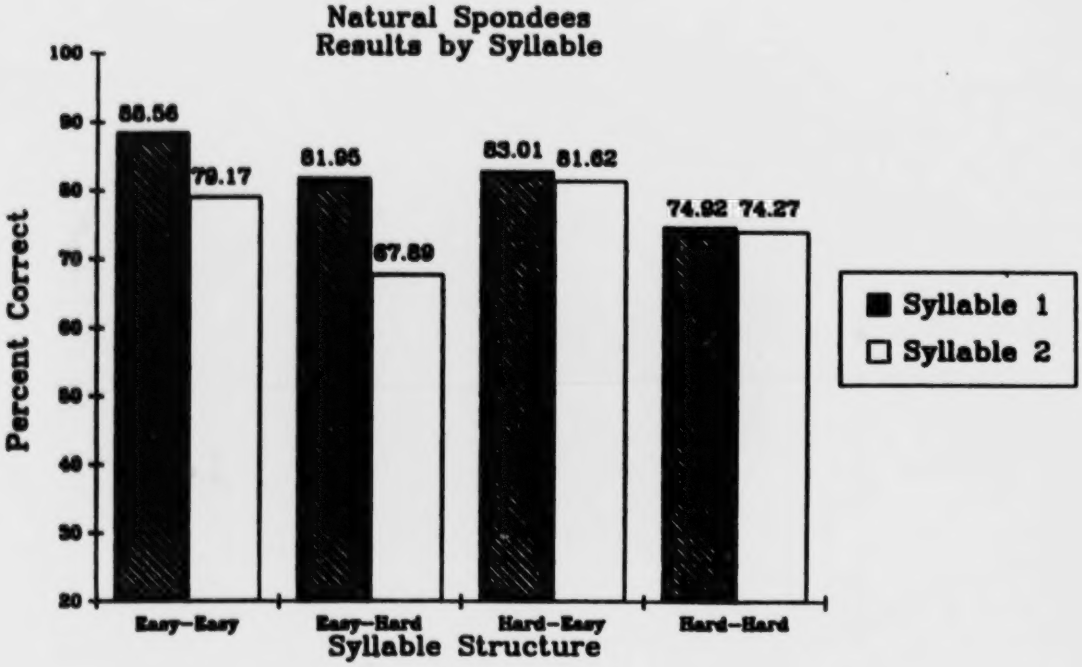


Figure 4. Percent correct identification for the individual syllables in the natural spondee condition.

the initial syllables are productive prefixes in English, the non-initial stressed syllables are not morphemes. That is, *-port* in *report* and *-scend* in *descend* do not constitute independent units of meaning in English, whereas *-type* in *retype* does. More to the point, according to Cutler's model, the weak initial syllables (prefixes) should not initiate lexical hypotheses or even segmentation. Although it is less clear at this time, it is possible to conjecture that the second syllable may act retroactively to help resolve whether the preceding phonetic information is included with the second syllable or whether it stands alone as a function word (e.g., *a* in *avoid* versus *a* *void*). Thus, although prefixes do not seem to initiate lexical hypotheses, non-initial stressed syllables may help to integrate the preceding syllable information in a retroactive manner.

However, Shillcock presents results suggesting that the system interprets syllabic information differently depending upon the success of obtaining semantic information. For example, *-bone* in the monomorphemic *trombone* primed *rib*, but *-port* in the pseudo-prefixed *report* did not prime *wine*. Shillcock argues that priming occurred in the former case because the second syllables in low-density, low-frequency neighborhoods act retroactively to facilitate integration of preceding syllabic information may be contingent upon the syllable also having morphemic status. Stated otherwise, retroactive integration of syllables may depend upon initial lexical hypotheses activating *words* (i.e., meaningful units) in the lexicon and therefore all linguistic information associated with them.

In conclusion, the findings from Shillcock, Cutler, and the spondee data suggest that non-word-initial strong syllables activate lexical hypotheses. Moreover, the spondee data suggest that non-initial syllables retroactively may influence the integration of preceding syllabic information. Specifically, when non-initial syllables are in low-frequency, low-density neighborhoods, they facilitate identification of initial syllables in high-frequency, high-density neighborhoods. As Shillcock and Cutler note for their own results, strict sequential models of lexical access cannot account for all aspects of lexical access in continuous speech. With regard to the spondee data, such models cannot account for the evidence showing retroactive influence of non-word-initial syllables in low-density, low-frequency neighborhoods helping to resolve initial syllables in high-density, high-frequency neighborhoods.

References

Cluff, M. S. & Luce, P. A. (1988). Neighborhoods of spondees in spoken word recognition. (In preparation.)

Cutler, A. (in press). In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Cutler, A. & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language*, **2**, 133-142.

Cutler, A. & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, **14**, 113-121.

Grosjean, F. & Gee, J. (1987). Prosodic structure and spoken word recognition. *Cognition*, **25**, 135-155.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6.* Bloomington, IN: Indiana University.

Luce, P. A., Pisoni, D. B., & Goldinger, S. D. (in press). Similarity neighborhoods in spoken words. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Shillcock, R. C. (in press). Lexical hypotheses in continuous speech. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 14 (1988)
Indiana University

Priming Lexical Neighbors of Spoken Words: Effects of Competition and Inhibition[1]

Stephen D. Goldinger, Paul A. Luce,[2] and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] **This research was supported by NIH Research Grant *NS-12179-12* to Indiana University in Bloomington, IN. We thank Jan Charles-Luce and Michael Cluff for their assistance in evaluating the stimuli, Richard Shiffrin, Gabriel Frommer, Denise Beike, Edward Shoben and two anonymous reviewers for comments and suggestions, and Michael Dedina for his technical advice.**

[2] **Paul Luce is now with the Department of Psychology, State University of New York at Buffalo.**

Abstract

Two experiments employing an auditory priming paradigm were conducted to test predictions of the Neighborhood Activation Model of spoken word recognition (Luce & Pisoni, in press). Acoustic-phonetic similarity, neighborhood densities, and frequencies of prime and target words were manipulated. In Experiment 1, priming with low frequency, phonetically related spoken words inhibited target recognition, as predicted by the Neighborhood Activation Model. In Experiment 2, the same prime-target pairs were presented with a longer inter-stimulus interval and the effects of priming were eliminated. In both experiments, predictions derived from the Neighborhood Activation Model regarding the effects of neighborhood density and word frequency were supported. The results are discussed in terms of competing activation of lexical neighbors and the dissociation of activation and frequency in spoken word recognition.

Priming Lexical Neighbors of Spoken Words: Effects of Competition and Inhibition

A fundamental problem in research on spoken word recognition concerns the processes by which stimulus information in the speech waveform is mapped onto lexical representations in long-term memory. Clearly, given the enormous size of the adult mental lexicon, isolating the sound pattern of one word from among tens of thousands of others in memory is no trivial matter for the listener. Nevertheless, word recognition normally appears to proceed effortlessly and with few errors. Given that one of the primary tasks of the word recognition system involves discrimination among lexical items, the study of the structural organization of words in memory takes on considerable importance for research in word recognition. In the present context, "structure" is defined specifically in terms of similarity relations among the sound patterns of words.

In a series of recent experiments, Luce and Pisoni (1988) investigated the effects that the number and nature of words activated in memory have on word recognition. Specifically, they examined the recognition of words in different *similarity neighborhoods*. A similarity neighborhood is defined as a collection of words that are phonetically similar to a given stimulus word. Two key structural characteristics been used to describe similarity neighborhoods. *Neighborhood density* refers to the absolute number of words occurring in any given similarity neighborhood; *neighborhood frequency* refers to the frequencies of occurrence of the neighbors.

In order to determine the effects of similarity neighborhood structure on spoken word recognition, Luce and Pisoni performed experiments employing three paradigms: perceptual identification, auditory lexical decision, and auditory word naming. In these studies, the density and frequency characteristics of similarity neighborhoods were found to be important determinants of the speed and accuracy of stimulus identification. In brief, the major results of their study were the following:

First, words that came from sparse neighborhoods, that is, neighborhoods that contain few other phonetically similar words, were recognized more quickly and more accurately than words that came from more dense neighborhoods. Second, words having primarily low frequency neighbors were recognized more quickly and more accurately than words having primarily high frequency neighbors. In addition to neighborhood frequency effects, Luce and Pisoni also found word frequency effects—high frequency words were identified better than low frequency words. This was the case in both the perceptual identification and the auditory lexical decision studies, but not in the auditory word naming study. In the naming experiment, reliable effects of neighborhood density as described above were observed, but the effects of both neighborhood frequency and item frequency were largely attenuated.

To account for these findings, Luce and Pisoni (1988) have proposed the Neighborhood Activation Model (NAM) of spoken word recognition. A flow chart of NAM is shown in

Figure 1. The model states that upon presentation of stimulus input, a set of acoustic-phonetic patterns is activated in memory. It is assumed that all patterns similar to the input are activated regardless of whether they correspond to real words in the lexicon or not. These acoustic-phonetic patterns then activate a system of *word decision units* tuned to the patterns themselves. Only those acoustic-phonetic patterns corresponding to words in memory will activate word decision units. Neighborhood activation is therefore assumed to be identical to the activation of the word decision units.

Insert Figure 1 about here

A diagram of a single word decision unit is shown in Figure 2.

Insert Figure 2 about here

Once activated, these decision units monitor the activation levels of the acoustic-phonetic patterns to which they correspond, as well as higher-level lexical information. Word frequency is included in the higher-level lexical information available to the decision units. These units therefore serve as the interface between acoustic-phonetic information in the speech waveform and higher-level lexical information in long-term memory. Acoustic-phonetic information is assumed to drive the system by activating the word decision units whereas higher-level lexical information is assumed to operate by biasing these decision units. These biases operate by adjusting the activation levels represented within the word decision units.

The values computed by the word decision units for determining whether a particular pattern has been presented are given by the Neighborhood Probability Rule, which has the form:

$$p(ID) = \frac{SWP * freq_s}{SWP * freq_s + \sum_{j=1}^{n}[NWP_j * freq_j]}$$

in which SWP is the probability of the stimulus word, $freq_s$ is the frequency of the stimulus word, NPW_j is the probability of neighbor j, and $freq_j$ is the frequency of neighbor j.

Neighborhood Activation Model (NAM)

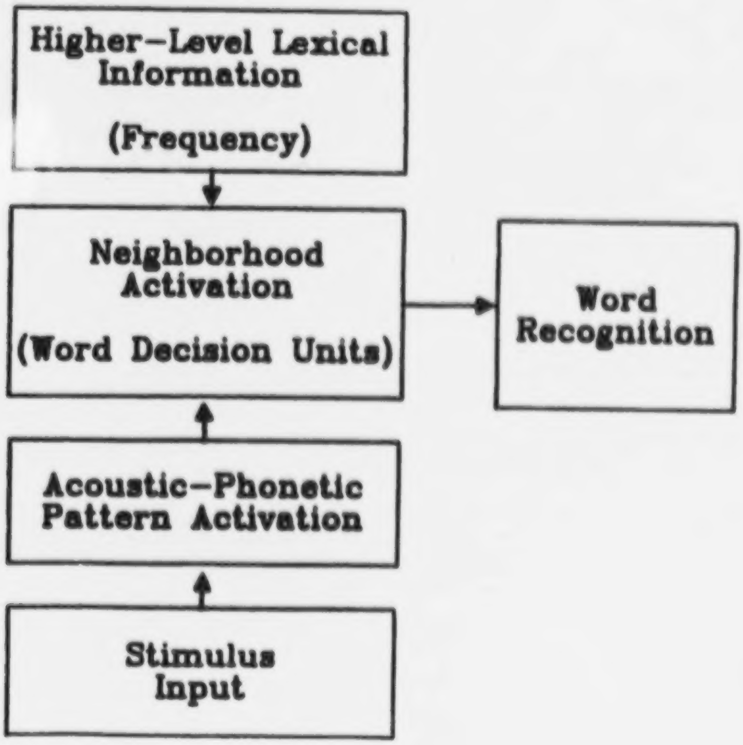

Higher-Level Lexical Information
(Frequency)
Neighborhood Activation
(Word Decision Units)
Word Recognition
Acoustic-Phonetic Pattern Activation
Stimulus Input

Figure 1. Flow diagram of the Neighborhood Activation Model (from Luce, 1986).

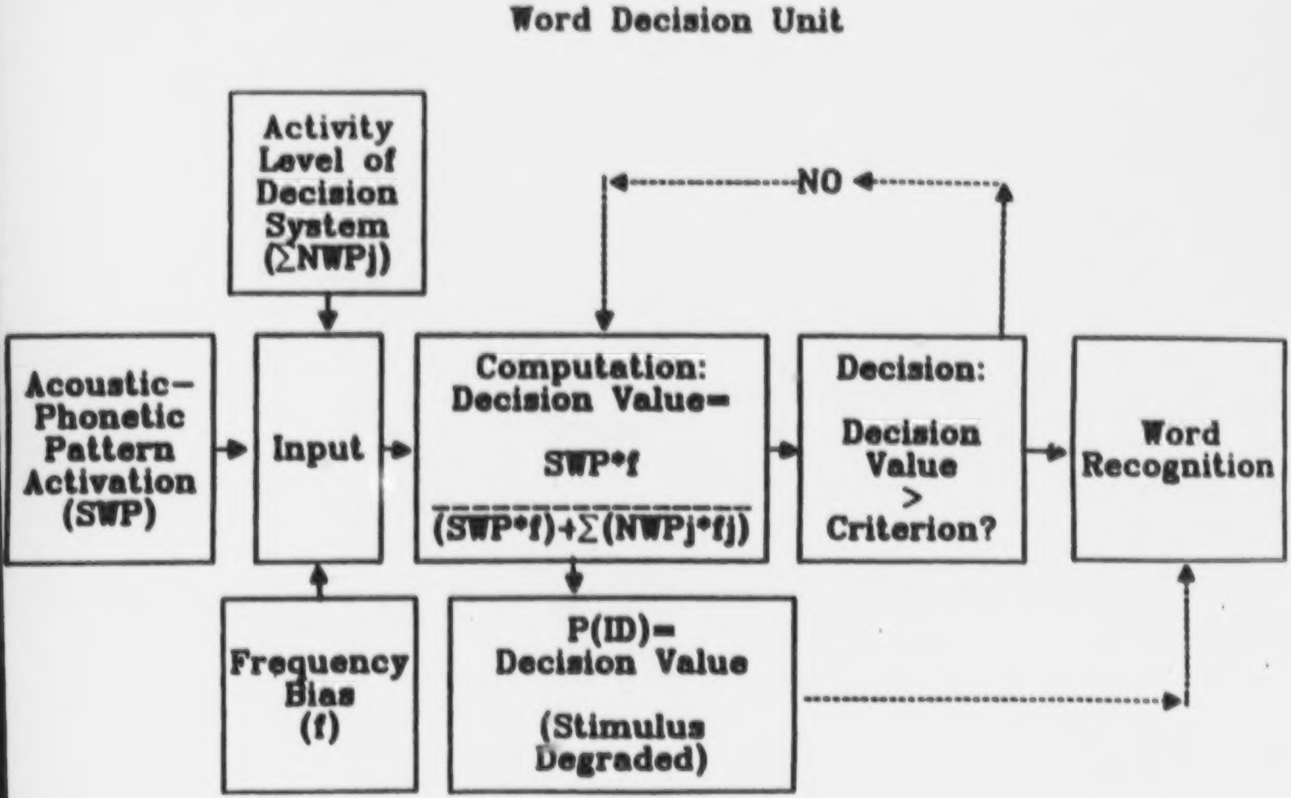

Word Decision Unit
Activity Level of Decision System (ΣNWPj)
NO
Acoustic-Phonetic Pattern Activation (SWP)
Input
Computation: Decision Value=
SWP*f / (SWP*f)+Σ(NWPj*fj)
Decision: Decision Value > Criterion?
Word Recognition
Frequency Bias (f)
P(ID)= Decision Value
(Stimulus Degraded)

Figure 2. Diagram of a single word decision unit (from Luce, 1986).

This rule, based on R.D. Luce's (1959) choice rule, combines neighborhood density, neighborhood frequency and stimulus word frequency to predict identification performance. The probability of identifying a stimulus word, $p(ID)$, is therefore equal to the probability of the stimulus word divided by the probability of the word plus the combined probabilities of its neighbors. Neighborhood density and neighborhood frequency are represented in the denominator term of the rule as the summed weighted probabilities of all neighbors. (In the present experiments, all values of "neighborhood density" for any given word actually refer to the frequency-weighted neighborhood probability for the word (see Luce, 1986). Throughout this paper, we will be using the general term "density" to denote this joint metric, and we will not discuss neighborhood frequency as an explicit variable.) Frequency is represented as a weighting function on the stimulus and neighbor probabilities, biasing decisions in favor of higher frequency words.

As shown in Figure 2, each decision unit is responsible for monitoring three sources of information that are simultaneously accounted for by the neighborhood probability rule: acoustic-phonetic pattern activation (SWP), higher-level lexical information ($freq_i$ and $freq_j$), and the overall level of activity in the system of units (the sum of the NPW_j's). As analysis of the stimulus input proceeds, the decision units continuously compute decision values via the neighborhood probability rule. As more information accumulates, the acoustic-phonetic pattern corresponding to the stimulus input is resolved. As the pattern is resolved, the activation levels of *similar* patterns steadily decrease and the decision values computed by the word decision unit monitoring the pattern of the actual stimulus steadily increase. Once the output of a given decision unit reaches some criterion, all information monitored by that decision unit is made available to working memory.

The neighborhood probability rule predicts reduced identification as a function of increased neighborhood density, which is represented in the denominator term of the rule. In terms of the on-line processing of the decision units, the presence of many similar neighbors serves to prolong the amount of processing time necessary to resolve the given input pattern. According to the model, frequency acts to bias decisions toward higher frequency items. Thus, the presence of high frequency words in the neighborhood is predicted by the rule to inhibit target identification. Note that NAM posits that activation in memory occurs independently of item frequencies and that the effects of frequency are not realized until the selection phase of the recognition process. Put another way, frequency is assumed to exert its influences *after* initial activation and *before* lexical access occurs. This claim regarding the role of frequency in word recognition is contrary to the claims of several other influential models of spoken word recognition, such as logogen theory (Morton, 1969), Forster's (1976) search model, and Marslen-Wilson's most recent version of cohort theory (1987).

Word recognition in NAM may be accomplished in a number of ways, depending on the requirements of the task. In situations in which the stimulus input is degraded, word recognition is accomplished by comparing the values computed by several word decision units and selecting the response corresponding to the highest value. When speeded responses are required, it is assumed that the subject sets a criterion for responding that, once exceeded by the output of a decision unit, results in the recognition of a word. Word recognition is defined explicitly as the choice of a particular pattern by the system of decision units.

One interesting prediction of NAM derives from the assumption of competition among lexical neighbors. As the neighborhood probability rule shows, increasing the activation level of a stimulus word's neighborhood is predicted to lower the probability of identifying the stimulus word itself. One means of experimentally manipulating the activation level of a stimulus word's neighborhood is to prime the stimulus word with one of its phonetically related neighbors. The model predicts that if a target stimulus presented for identification is immediately preceded by a phonetically related prime (a neighbor), residual activation from the prime should produce increased competition from the neighborhood, and thereby inhibit stimulus word identification. In terms of the neighborhood probability rule, priming with a phonetically related word should increase the $\sum[NPW_j * freq_j]$ term in the denominator of the rule and, therefore, reduce predicted identification performance for the target. In short, the model predicts inhibition priming from phonetically similar words.

In order to test this prediction, prime-target pairs related only by phonetic similarity were generated. In addition, as a baseline against which to evaluate predicted effects of inhibition priming, phonetically unrelated prime-target pairs were also generated. Thus, primes were either phonetically related or unrelated to the target words. None of the prime words were semantically related to the target words. Examples of target-related prime-unrelated prime sets are: *VEER-BULL-GUM*, *PAR-TALL-BASE*, and *HASH-ETCH-LAME*. [1]

Phonetic similarity between primes and targets was determined from confusion matrices for individual consonants and vowels obtained in a previous study (see Luce & Pisoni, 1988). Based on these confusion matrices, primes were chosen that constituted the nearest neighbors of the target words, with the constraint that the primes and targets shared no common phonemes. The restriction against overlapping phonemes was imposed in order to prevent subjects from generating response strategies based on repeated overlap between prime-target pairs (for discussion, see Slowiaczek, Nusbaum & Pisoni, 1987). [2]

---

[1] The authors may be contacted for complete lists of the stimuli used.

[2] Phonetic similarities between primes and targets were computed by determining the values of *SWP* and *NWP* in the neigborhood probability rule. The actual values of *SWP* and *NWP* in the neighborhood probability rule were computed as follows: Individual confusion matrices for all initial and final consonants and vowels were obtained under appropriate S/N ratios. The stimulus word probabilities (*SWPs*) were then computed by multiplicatively combining the probabilities of the initial consonants, vowels and final consonants of the target stimulus words to render estimates of the *SWPs* based on the confusion matrices. For example, for the stimulus word /kot/ ("coat"), the stimulus word probability was computed as follows: *SWP*(/kot/)= p(k|k)*p(o|o)*p(t|t). This product expresses the probability of the /k/ in /kot/ given that /k/ was actually presented, the probability of /o/ given /o/ was actually presented, and the probability of /t/ given /t/ was actually presented. Using the confusion matrices in this manner, it was thus possible to obtain an estimate of p(/kot/|/kot/). In this manner, *SWPs* for all the target words were computed.

In addition to the manipulation of prime type (related vs. unrelated), three other variables were examined: neighborhood density, prime frequency, and target frequency. Prime-target pairs were selected from dense, high frequency neighborhoods or from sparse, low frequency neighborhoods. Neighborhood density was manipulated, in part, to replicate previous work and, in part, to determine if the predicted inhibition priming would be influenced by the structure of the neighborhoods from which the prime-target pairs were drawn.

The frequencies of the primes and targets were manipulated by orthogonally combining two levels of prime frequency (high and low) with two levels of target frequency (high and low). This manipulation was included as an important test of one aspect of NAM, namely, the assumption that frequency affects decision processes and is not directly coded in the activation levels of the word patterns. This assumption of the model directly follows from the results of the auditory word naming experiment discussed above. In the naming study, Luce and Pisoni found that while neighborhood density remained a powerful predictor of performance, word frequency did not. Similarly, Balota and Chumbley (1984) showed that word frequency effects may be greatly attenuated in certain experimental settings. Certainly, if word frequency information were so deeply ingrained in the activation coding of words, such effects should not be possible.

Assuming for the moment, however, that frequency *directly* modifies activation levels, one could argue that high frequency primes should produce relatively more inhibition of target recognition than low frequency primes, simply because high frequency primes should produce stronger competing activation levels in the neighborhood. However, this result is *not* predicted by NAM. Because NAM states that frequency does not directly affect activation levels, high frequency primes should not produce any more activation than low frequency primes. Thus, NAM does not predict substantial inhibition of target recognition for targets preceded by high frequency primes.

However, NAM does predict differential priming effects as a function of the frequency of the prime. In fact, the model predicts, somewhat counter-intuitively, that *low* frequency primes should produce more inhibition than high frequency primes. The rationale for this prediction is as follows: All things being equal, NAM predicts that low frequency words should be identified less quickly and less accurately than high frequency words. Recall that this prediction is *not* based on the assumption that high frequency words have higher resting activation levels, lower recognition thresholds, or steeper activation functions than low

---

In order to obtain estimates of the *NWPs*, a similar procedure was employed. For example, in order to determine the *NWP* for /bæt/ ("bat") given presentation of the stimulus word /kot/, the confusion matrices were once again consulted. However, in this instance, an estimate of p(/bæt/|/kot/) was computed by multiplicatively combining the p(b|k), p(æ|o) and p(t|t). Thus, in the present study, those neighbors having the highest values for *NWP* that had no phonemic overlap with the target items were selected as primes.

frequency words. Instead, the word frequency advantage is assumed to arise because biased decisions regarding the stimulus input can be made more quickly and accurately for high frequency words. Therefore, activation levels for acoustic-phonetic patterns corresponding to high and low frequency words are assumed to rise and fall at the same rates. However, it is assumed that decisions in the word decision units can be made earlier for high frequency words than for low frequency words. If this is the case, the word decision unit for a high frequency word will surpass criterion for recognition sooner than a word decision unit for a low frequency word. This means that, in turn, the activation of an acoustic-phonetic pattern corresponding to a high frequency prime will begin to return to a resting level sooner than the activation of an acoustic-phonetic pattern corresponding to a low frequency prime. Thus, target items following high frequency primes should receive less competition from the residual activation of the prime than targets following low frequency primes.

In short, two main predictions were examined: First, it was predicted that phonetically related primes would inhibit target identification because of increased neighborhood competition. Second, it was predicted that, because frequency is assumed to affect decision processes and not activation levels, low frequency primes would produce relatively more inhibition than high frequency primes. This prediction is in contrast to predictions of several current models, discussed in detail below, that assume that frequency directly affects activation levels.

Experiment 1A

Method

Subjects. Sixty Indiana University undergraduate students participated in partial fulfillment of requirements of an introductory psychology course. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimuli. Two-hundred and forty phonetically related prime-target pairs were selected from a computerized lexical database based on Webster's pocket dictionary (1967). In addition, unrelated primes were selected for each of the 240 targets, for a total of 720 words. The related prime-target pairings were created by searching the database for each target's nearest neighbor with no common phonemes (see footnote 2). As stated above, degree of similarity of a given prime to its target word was computed using confusion matrices for individual consonants and vowels (see Luce, 1986, for a complete description). The unrelated primes were selected by searching for words from neighborhoods that had the same density as their prospective targets, but were not phonetically confusable with the targets. From the original lists of words generated by these searches, the final 720 words selected were those which met the following constraints: (1) All targets and unrelated primes were three phonemes in length; related primes were either two or three phonemes in length; (2) all words were monosyllabic; (3) all words were listed in the Kučera and Francis (1967) corpus; and (4) all words had a rated familiarity of 6.0 or above on a seven-point scale. These familiarity ratings were obtained from a previous study by Nusbaum, Pisoni, and Davis (1984). In this study, all words from Webster's pocket dictionary were presented visually to subjects for familiarity ratings. The rating scale ranged from (1) "don't know the word" to (4) "recognize the word but don't know its meaning" to (7) "know the word and its meaning." The rating criterion of 6.0 and above was used to ensure that all prime and target words would be known by the subjects.

After all constraints were satisfied, the sets of primes and targets were divided into eight cells constructed by orthogonally combining two levels (high and low) of each of four variables: 1) prime-target relatedness, 2) neighborhood density, 3) prime frequency and 4) target frequency. Once the prime-target pairs were assigned to their proper cells, the cell with the fewest pairs contained 30 items. Prime-target pairs with the lowest estimated phonetic confusability were removed from all other cells to leave 30 prime-target pairs in each of the eight cells, representing a virtually exhaustive set of all possible stimuli for the purposes of this experiment.

Once all the excess stimuli had been eliminated from the set of possible stimuli, the frequencies of the remaining words were as as follows: Low frequency targets ranged in log frequency from 1.4241 to 1.5378, with a mean log frequency of 1.4891; high frequency targets ranged in log frequency from 2.8204 to 3.1207, with a mean log frequency of 2.9057. Low frequency related primes ranged in log frequency from 1.4365 to 1.5984 with a mean log frequency of 1.5102; high frequency related primes ranged in log frequency from 2.6972 to 3.0346 with a mean log frequency of 2.9685. Low frequency unrelated primes ranged in log frequency from 1.5012 to 1.8441, with a mean log frequency of 1.5990; high frequency unrelated primes ranged in log frequency from 2.1025 to 2.9361, with a mean log frequency of 2.6408.

Because every target item had two corresponding primes and no subject was to be presented the same target item twice, the stimuli were divided into two lists. Every subject responded to all 240 targets, but the primes and control items varied. For a given group, 120 of the targets were primed by related primes, the other 120 targets by control primes. The next group received the same targets paired with their primes in the reverse order. An equal number of subjects were presented with each list.

The stimuli were recorded in a sound-attenuated booth by a male talker of a midwestern dialect using an Ampex AG500 tape deck and an Electro-Voice D054 microphone. All words were spoken in isolation. The stimuli were then low-pass filtered at 4.8 kHz and digitized at a sampling rate of 10 kHz using a 12-bit analog-to-digital converter. All words were excised from the list using a digitally controlled speech waveform editor (WAVES) on a PDP 11/34 computer (Luce and Carrell, 1981). The mean duration of the targets was 691.99 msec; mean durations for the related and unrelated primes were 705.29 and 699.84 msec, respectively. Finally, all words were paired with their appropriate counterparts and stored digitally as stimulus files on a computer disk for later real-time presentation to subjects during the experiment.

To ensure that all stimuli could be identified accurately, 10 additional subjects were asked to identify all words in the absence of noise. Words which were not correctly identified by at least 8 of 10 subjects were re-recorded and replaced with more intelligible tokens.

Design. Two levels of four variables were examined: (1) prime type (related vs. unrelated); (2) neighborhood density (high vs. low); (3) prime frequency (high vs. low); and (4) target frequency (high vs. low). The dependent measure in all cases was the percentage of target words correctly identified.

Procedure. Subjects were tested in groups of five or fewer. Each subject was seated in a testing booth equipped with an ADM computer terminal and a pair of TDH-39 headphones. The presentation of stimuli was controlled by a PDP 11/34 computer. All stimuli were presented in random order.

A typical trial proceeded as follows: A prompt would appear on the CRT screen saying, "GET READY FOR NEXT TRIAL." Five hundred msec after this prompt appeared, a prime was presented over headphones at 75 dB (SPL) in the clear. Immediately upon the offset of the prime, 70 dB (SPL) of continuous white noise was presented. Fifty msec after presentation of the noise, the target item was presented at 75 dB (SPL), yielding a +5 dB signal-to-noise (S/N) ratio. The subjects' task was to identify each target word and type their responses on the ADM keyboard as accurately as possible following each trial. Subjects were under no time constraints to respond.

Each subject performed 280 trials, the first 40 of which were practice and were not included in the final data analysis. There were equal numbers of trials in all 16 conditions, so each subject identified a target from each condition 15 times. Across 60 subjects, this procedure generated a total of 900 responses per condition.

Results

The percentage of words correctly identified was determined for each subject. For a response to be considered correct, the entire response had to match the target item exactly or be a homophone (e.g., *there, their*). Responses were corrected for simple spelling errors prior to analysis.

Figure 3 displays the results of the priming manipulation for all conditions. Light bars indicate conditions for unrelated primes, dark bars show performance for related primes. Mean percentage of correct target identification for high frequency targets is shown on the left; performance for low frequency targets is shown on the right. Performance for prime-target pairs selected from sparse neighborhoods is shown in the upper panel; performance for prime-target pairs from dense neighborhoods is shown in the lower panel.

Insert Figure 3 about here

A four-way analysis of variance (prime type X neighborhood density X prime frequency X target frequency) was performed on the mean percentages of correct responses. A significant main effect of prime type was obtained [$F(1,59)=8.11$, $MS_e=.0146$, $p < .05$]. (All results reported are $p < .05$ or beyond, unless specifically stated otherwise). Post-hoc Tukey's HSD analyses indicated that targets following related primes were identified significantly less accurately than targets following unrelated primes in three conditions. These conditions, denoted by asterisks in Figure 3, are: (1) dense neighborhood/low frequency prime/high frequency target; (2) sparse neighborhood/low frequency prime/high frequency target; and (3) sparse neighborhood/low frequency prime/low frequency target. Thus, significant inhibition was obtained only when targets were preceded by low frequency primes.

The effects of neighborhood structure and target frequency are shown in Figure 4. These results are collapsed across prime type to better illustrate the effects of neighborhood density and target frequency. Light bars indicate targets from sparse neighborhoods; dark bars indicate targets from dense neighborhoods. Mean percentage of correct target identification for high frequency targets is shown on the left, whereas performance for low frequency targets is shown on the right.

Insert Figure 4 about here

A significant main effect of neighborhood density was obtained [$F(1,59)=248.28$, $MS_e = .0161$]. In all conditions, target words occurring in sparse neighborhoods were recognized

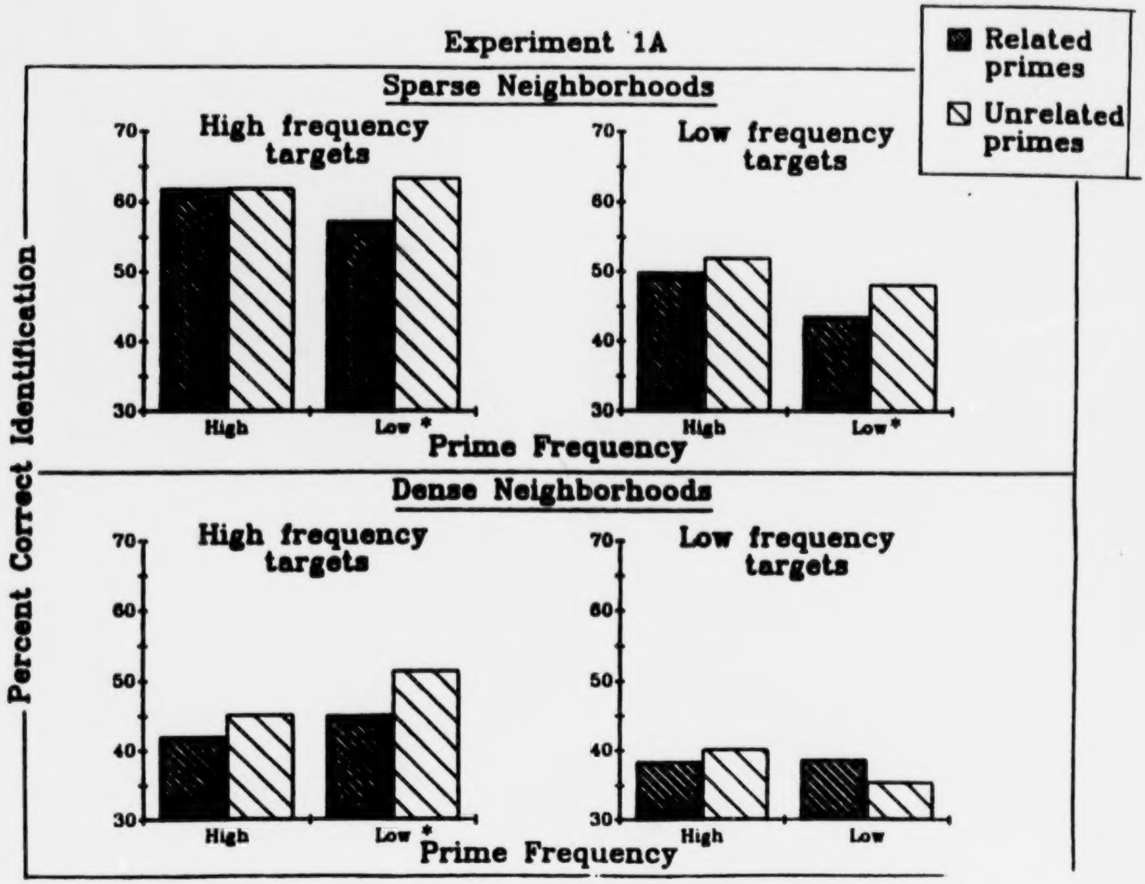


Figure 3. Percent correct identification for high and low frequency target words as a function of neighborhood density and prime frequency for related and unrelated primes. Light bars show performance for neutral primes, dark bars show performance for related primes.

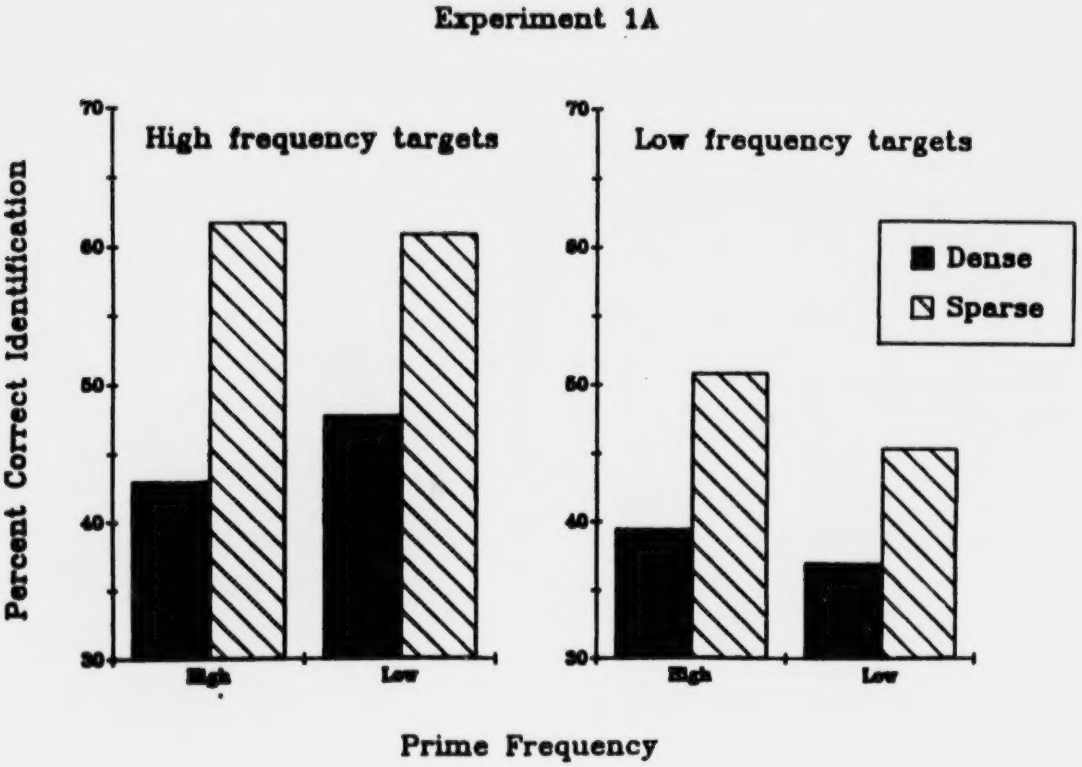


Figure 4. Percent correct identification for high and low frequency target words as a function of neighborhood density and prime frequency, averaged over prime type. Light bars show performance for targets from sparse neighborhoods, whereas dark bars show performance for targets from dense neighborhoods. The left panel shows the results for high frequency targets: the right panel shows the results for low frequency targets.

more accurately than target words occurring in dense neighborhoods. A significant main effect of target frequency was also obtained [$F(1,59)=164.00$, $MS_e=.0158$]. In all conditions, high frequency targets were recognized more accurately than low frequency targets. There was no main effect of prime frequency [$F(1,59)=2.35$, $MS_e=.0091$, $p=.1306$].

The ANOVA based on subject performance showed several significant interactions. There were interactions of neighborhood density X target frequency [$F(1,59)=18.23$, $MS_e=.0098$], of neighborhood density X prime frequency [$F(1,59)=12.17$, $MS_e=.0088$], and of target frequency X prime frequency [$F(1,59)=21.22$, $MS_e=.0087$].

In addition to the ANOVA performed on the data grouped by subjects, an item analysis was performed to make sure that the results were not caused by idiosyncratic stimuli randomly assigned to certain cells in the design (Clark, 1973). This analysis was performed even though we considered our stimuli an exhaustive set of all possible pairs conforming to the constraints enumerated above. All the main effects obtained above were also obtained in the item analysis. Main effects were obtained for prime type [$F(1,232)=5.06$, $MS_e=.0709$], for neighborhood density [$F(1,232)=12.14$, $MS_e=.1660$], and for target frequency [$F=7.82$, $MS_e=.1660$]. No main effect of prime frequency was obtained in the item analysis [$F(1,232)=0.14$, $MS_e=.1660$, $p=.7120$] nor were any of the interactions obtained in the ANOVA performed on the subject data significant in the item analysis.

The basic pattern of results from this experiment is clear. In three of eight experimental conditions, priming with a phonetically related word significantly *inhibited* target recognition. A strong trend toward inhibition was also observed in four of the five remaining conditions. Furthermore, significant effects of neighborhood density and target frequency were observed across all conditions. Words occurring in sparse neighborhoods were identified more accurately than words occurring in dense neighborhoods. Also, high frequency words were identified more accurately than low frequency words. All three of these findings are consistent with the predictions outlined in the introduction. The results of the neighborhood density and frequency manipulations are consistent with the earlier findings of Luce and Pisoni (1988).

The three individual conditions in which significant inhibition effects were obtained shared a common property: All three conditions contained low frequency primes. Although there was no overall main effect of prime frequency, these findings demonstrate that low frequency primes do indeed inhibit target recognition more than high frequency primes, as predicted by NAM.

The inhibition demonstrated in Experiment 1A presumably was caused by increased competition in the target words' neighborhoods due to the lingering activation of the prime. When a target is presented only 50 msec after a low frequency prime, it is assumed that *more* residual activation is still in the neighborhood than when a target follows a high frequency prime. This occurs because decisions regarding low frequency words are made more slowly than decisions regarding high frequency words, thus allowing activation levels for low frequency primes to linger for more time. As a consequence, more lexical candidates remain activated when the target is presented, thereby producing greater competition between the target and possible alternatives.

Experiment 1B

Although the results indicating inhibition priming obtained in Experiment 1A were statistically significant under both subject and item analyses, a replication of these findings was undertaken because of the potential theoretical importance of the present set of results. An exact replication of Experiment 1A was performed with 38 new subjects. The stimuli, experimental procedure, subject characteristics and data analyses were identical to those used in Experiment 1A.

Results

Figure 5 displays the results of the priming manipulation for all conditions. As in Experiment 1A, a four-way analysis of variance (prime type X neighborhood density X prime frequency X target frequency) was performed on the mean percentages of correct responses.

Insert Figure 5 about here

A significant effect of prime type was again obtained [$F(1,37)=36.10$, $MS_e=.0081$, $p <$ .05]. (All reported results are significant at the .05 level or beyond.) In addition, a significant main effect of prime frequency was obtained [$F(1,37)=4.27$, $MS_e=0.1193$], such that low frequency primes produced more inhibition than high frequency primes. As in Experiment 1A, item analyses were also conducted to ensure that the effects were not due to any idiosyncratic stimulus items. By the item analysis, main effects of prime type [$F(1,224)=8.51$, $MS_e=.0375$] and of prime frequency [$F(1,224)=8.43$, $MS_e=.0357$] were again obtained. Therefore, not only was the effect of inhibition priming significant in the replication, but the overall effect of prime frequency, which was not significant in Experiment 1A, was now found to be statistically reliable in the replication. These findings provide strong support to the validity of the results obtained in Experiment 1A.

The predictions for Experiments 1A and 1B were based on the assumption that any priming effects obtained would arise purely from the increased amount of general activation

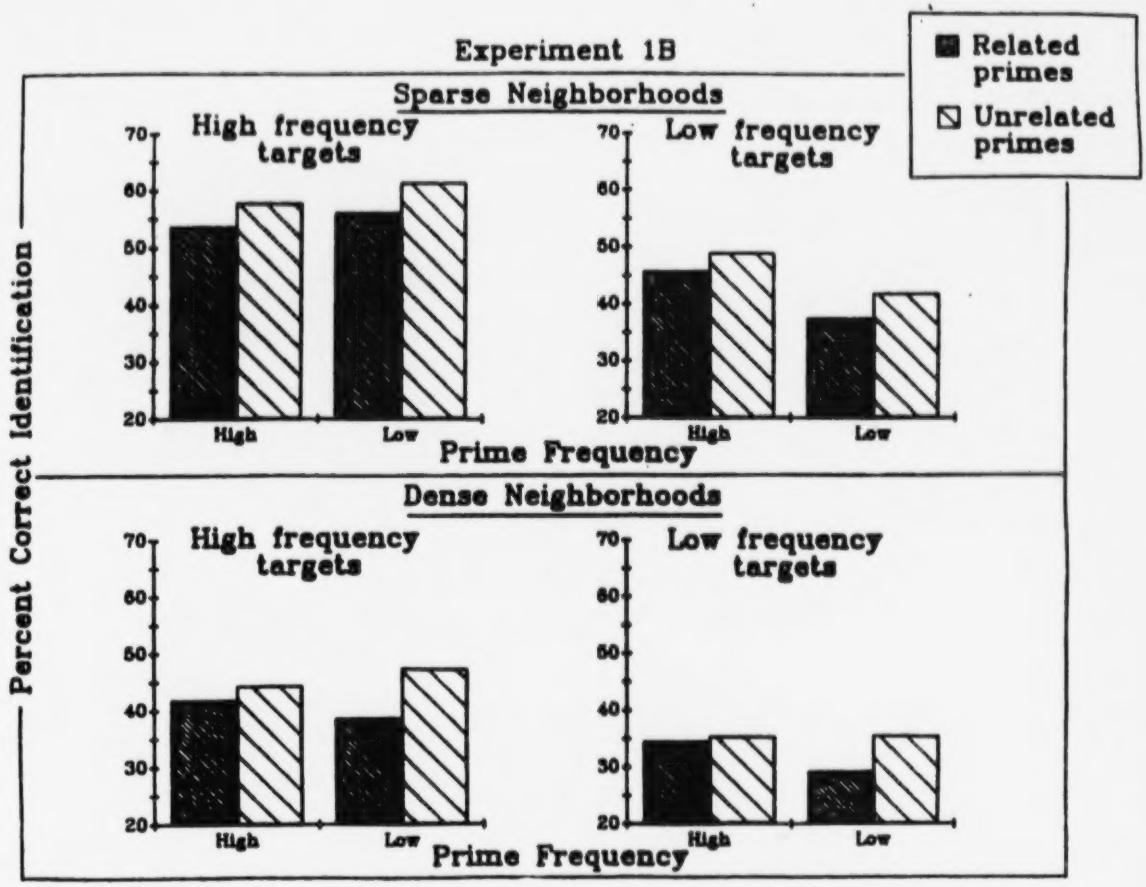


Figure 5. Percent correct identification for high and low frequency target words as a function of neighborhood density and prime frequency for related and unrelated primes. Light bars show performance for neutral primes, dark bars show performance for related primes.

in the primed similarity neighborhood. Incorporated into this assumption was the previously stated assumption that the primes and targets were related so subtly that subjects would not be able to apply any decision strategies due to learning during the course of the experiment. Experiment 2 was conducted to validate these assumptions and to further examine the time course of the observed priming effect. Experiments 1A and 1B employed a 50 msec delay between primes and targets. Experiment 2 was a replication of the first two experiments; however, the inter-stimulus interval was increased to 500 msec.

It is generally accepted that following stimulus recognition, activation in memory drops, returning eventually to some resting level (Collins & Loftus, 1975;, McClelland & Elman, 1986). If the priming effects demonstrated in the first experiment were due only to competition at an activation level, it should be possible to eliminate the inhibition effects simply by allowing the activation to fade over time. We predicted that increasing the inter-stimulus interval from 50 to 500 msec would eliminate the inhibition from priming observed in Experiments 1A and 1B. We also predicted, however, that the pattern of results found in Experiments 1A and 1B with regard to the effects of target frequencies and neighborhood densities would remain unchanged in Experiment 2.

Experiment 2

Method

The stimuli and experimental design were the same as Experiments 1A and 1B. The procedure was identical to that of Experiments 1A and 1B, with the exception of the increased inter-stimulus interval between prime and target items.

Subjects. Sixty different subjects participated in partial fulfillment of the requirements of an introductory psychology course. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Results

The percentage of words correctly identified was determined for each subject. As in Experiments 1A and 1B, for a response to be considered correct, the entire response had to match the target exactly or be a homophone (e.g., *there, their*). Responses were corrected for simple spelling errors prior to analysis.

Figure 6 displays the results of the priming manipulation for all conditions. The effects of neighborhood structure and target frequency are shown in Figure 7. These results are collapsed across prime type to better illustrate the effects of neighborhood density and target frequency.

Insert Figure 6 about here

Insert Figure 7 about here

A four-way analysis of variance (prime type X density X target frequency X prime frequency) was performed on the mean percentages of correct responses for each condition. Significant main effects were obtained for neighborhood density [$F(1,59)=435.53$, $MS_e=.0123$, $p < .05$] and target frequency [$F(1, 59)=245.58$, $MS_e=.0122$, $p < .05$]. (Again, as in Experiment 1A, all results are $p < .05$, unless stated otherwise). As in Experiments 1A and 1B, these main effects showed that target recognition was more accurate for words selected from sparse neighborhoods than for words selected from dense neighborhoods, and that target recognition was more accurate for high frequency targets than for low frequency targets. No main effects of prime type [$F(1,59)=1.45$, $MS_e=.0149$, $p=.233$] or prime frequency [$F(1,59)=3.01$, $MS_e=.012$, $p=.088$] were observed. However, a significant interaction of target frequency X neighborhood density was observed [$F(1,59)=6.30$, $MS_e=.0133$].

In addition to the overall ANOVA, an item analysis was again performed to ensure that the results were not due to a few idiosyncratic stimulus items. Significant main effects were again obtained for neighborhood density [$F(1,232)=18.45$, $MS_e=.1475$] and target frequency [$F(1,232)=10.07$, $MS_e=.1475$]. No main effects were obtained for prime

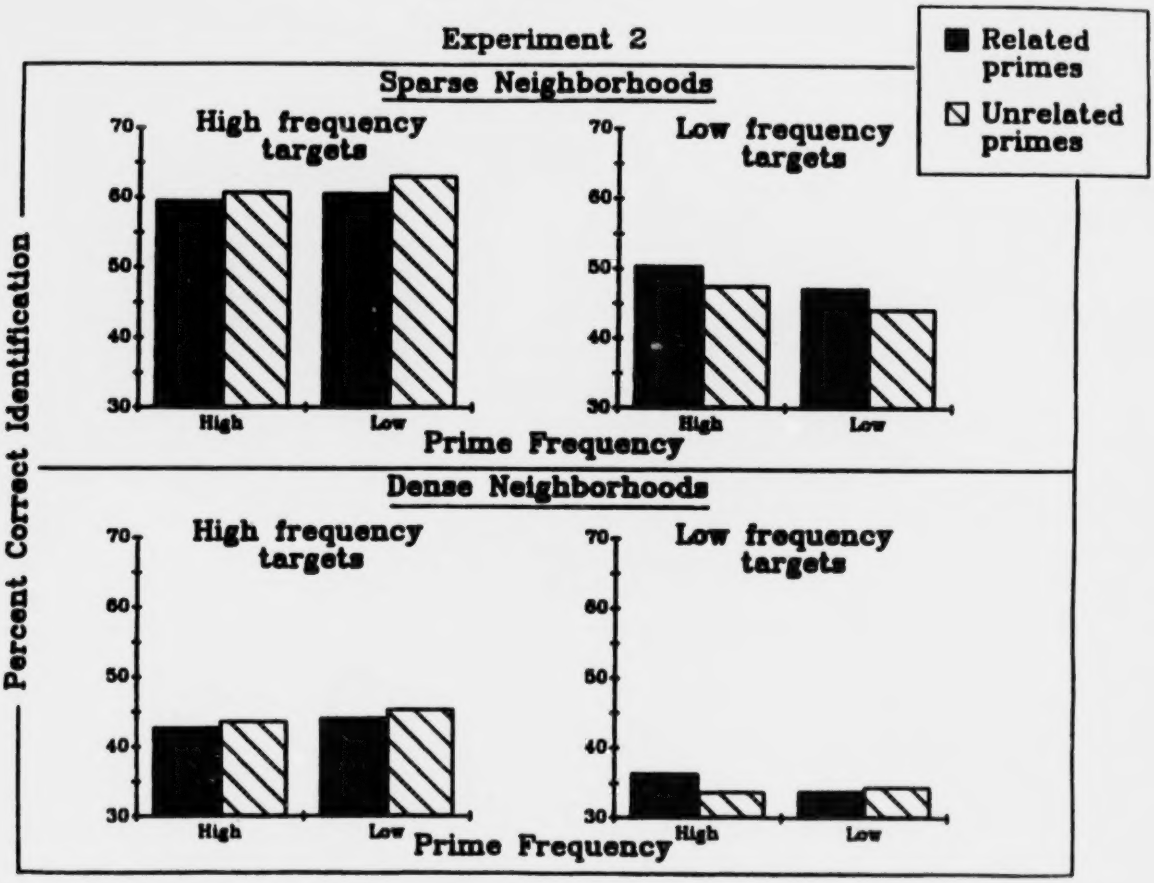


Figure 6. Percent correct identification for high and low frequency target words as a function of neighborhood density and prime frequency for related and unrelated primes. Light bars show performance for neutral primes; dark bars show performance for related primes.

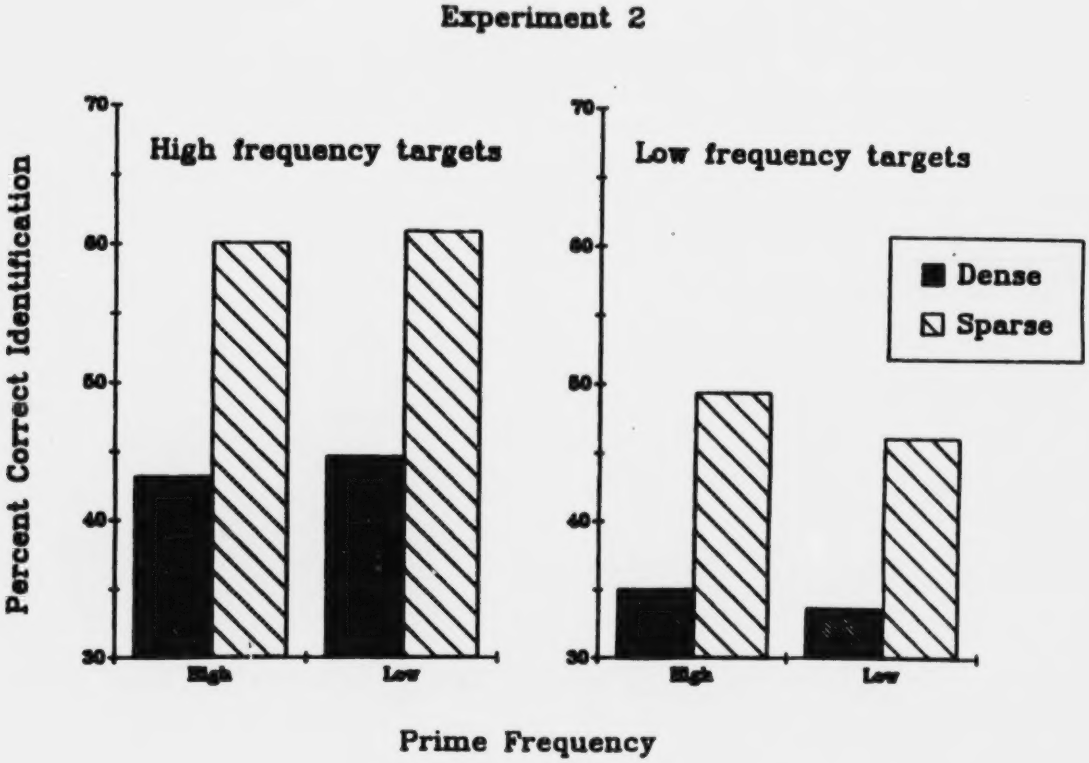


Figure 7. Percent correct identification for high and low frequency target words as a function of neighborhood density and prime frequency, averaged over prime type. Light bars show performance for targets from sparse neighborhoods, whereas dark bars show performance for targets from dense neighborhoods. The left panel shows the results for high frequency targets; the right panel shows the results for low frequency targets.

type [$F(1,232)=0.16$, $MS_e=.0126$, $p=.6864$] or prime frequency [$F(1,232)=.02$, $MS_e=.1475$, $p=.8783$]. There were no significant interactions obtained in the item analysis.

The results of Experiment 2 closely resemble those of Experiments 1A and 1B with respect to effects of neighborhood densities and target frequencies on target recognition. However, as expected, we did not find any effects of inhibition as observed in Experiments 1A and 1B. (Indeed, an analysis of variance performed across Experiments 1A and 2 revealed a significant interaction of prime type X inter-stimulus interval [$\mathrm{F}(1,118)=6.80$, $MS_e=.0124$, $p < .05$]). Our hypothesis was that the inhibition obtained in Experiments 1A and 1B was caused by increased activation of words in the target's neighborhood resulting from previous recognition of the prime item. Following this reasoning, we predicted that this inhibition would not occur when a longer inter-stimulus interval was used because the residual activation in the neighborhood produced by the prime would have more time to drop to its resting level. The results obtained supported these predictions and lend credence to our interpretation of Experiments 1A and 1B.

General Discussion

The present experiments were undertaken in order to test several predictions of the Neighborhood Activation Model using a priming paradigm. We were specifically interested in directly testing the Neighborhood Probability Rule. Recall that the rule states that as activation of a neighborhood increases, the probability of recognizing a given stimulus word in that neighborhood will decrease. Exploiting this property of the rule, we predicted that priming a lexical neighborhood with a word phonetically similar to a subsequent target word would inhibit target recognition. Indeed, in Experiments 1A and 1B we found that priming with a phonetically related neighbor inhibited target recognition, relative to a baseline determined by priming with unrelated words. The short-lived effect of inhibition from priming observed in Experiments 1A and 1B appears to arise purely from the competing activation among phonetically similar lexical neighbors. This is evident by the null result of Experiment 2, in which the effects of priming were eliminated entirely by increasing the inter-stimulus interval from 50 msec to 500 msec.

In addition to the effects of priming, we also found that both neighborhood densities and item frequencies influenced target identification accuracy. Target words from sparse neighborhoods were identified better than target words from dense neighborhoods, and high frequency target words were identified better than low frequency target words.

The present experiments replicate the earlier findings of Luce and Pisoni (1988). Using perceptual identification, lexical decision, and naming paradigms, Luce and Pisoni obtained consistent results indicating that neighborhood density, neighborhood frequency, and item frequency are primary determinants of spoken word recognition performance. The present study, employing an auditory priming paradigm, has again demonstrated reliable effects of these same structural properties of similarity neighborhoods. When subjects were presented with low frequency targets from dense neighborhoods, performance was worst, whereas when they were presented with high frequency targets from sparse neighborhoods, performance was best. The inhibition effects, as well as the neighborhood effects observed here were all predicted by NAM.

Most contemporary models of word recognition, such as logogen theory (Morton, 1969), Forster's search theory (Forster, 1976), and cohort theory (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1987) assume that word frequency is directly represented in the resting activation levels, or in the relative speed of activation of words in long-term memory. In contrast, NAM does not make this assumption. Rather, in NAM, decision units that are sensitive to frequency information operate after activation has occurred and bias the decision units toward higher frequency words. In the priming paradigm, the effect of the prime on the identification of the target item is generally believed to arise from residual activation in long-term memory (e.g., Collins & Loftus, 1975; McClelland & Rumelhart, 1981; Slowiaczek, Nusbaum & Pisoni, 1987). If activation arises directly from presentation of the prime, and if frequency information *directly* affects the level of activation, one would expect to find a reliable trend for *high* frequency primes to exert the largest influences on target identification. In the present experiment, we found that *low* frequency primes produce greater inhibiton of target recognition than high frequency primes, a result that may be problematic for several current models that assert that frequency is coded in the activation levels themselves. In contrast, we predicted that low frequency primes would be recognized more slowly than high frequency primes. Consequently, low frequency primes would leave more residual activation and competition in the neighborhood than high frequency primes. This unresolved activation would produce added competition among phonetically similar words, making low frequency primes inhibit subsequent target items more than high frequency primes. This prediction was, in fact, confirmed in Experiments 1A and 1B.

If word frequency were directly represented in the activation levels of words, however, high frequency primes should have produced stronger inhibition effects. This was clearly not the case, thus lending additional support to NAM's characterization of the role of frequency information in the decision stage of spoken word recognition.

In contrast to the present results demonstrating *inhibition* priming, numerous examples of *facilitation* priming in many different kinds of studies and different modalities of presentation have been reported in the literature (e.g., Collins & Loftus, 1980; Jakimik, Cole & Rudnicky, 1985; Slowiaczek, Nusbaum & Pisoni, 1987). All of these studies obtained facilitation from priming whether presentation was visual or auditory and whether the relations between primes and targets were semantic or phonological. In one recent experiment, Slowiaczek, Nusbaum, and Pisoni (1987) presented prime-target pairs which overlapped by zero, one, two, or three phonemes. Their results showed that accuracy of target identification improved significantly as the number of overlapping phonemes between primes and targets increased. The contrasting results of the Slowiaczek et al. study and the present study raise two important questions. First, we must ask whether or not the current finding of inhibition from priming is supported in the literature. Second, we must ask why some priming procedures produce inhibition while others produce facilitation.

Just as one can find numerous examples of priming studies resulting in facilitation of target identification, one can also cite examples of priming studies demonstrating inhibition. The majority of these findings have occurred in semantic priming experiments using visual presentation (e.g., Meyer, Schvaneveldt & Ruddy, 1974; Neely, Schmidt & Roediger, 1983; Taraban & McClelland, 1987), but there have been some findings in auditory priming (e.g., Tanenhaus, Flanigan & Seidenberg, 1980; Slowiaczek & Pisoni, 1986). Especially interesting are the recent findings of Taraban & McClelland (1987). These researchers used a primed naming paradigm and measured latency of onset to pronounce visually presented target words and obtained results that were similar to the results of the present study. The authors discuss "conspiracy models" of word pronunciation and note that visually presented words with many orthographically similar neighbors are pronounced more slowly than words with few similar neighbors. This effect is analogous to the effects of neighborhood density found by Luce & Pisoni (1988) and replicated in the present study of spoken word recognition. Furthermore, Taraban & McClelland found that visually similar words presented as prime-target pairs produced longer latencies to pronounce and more mispronunciations than dissimilar words used as prime-target pairs. This effect is analogous to the effect of inhibition from priming obtained in the present study. In another study, Slowiaczek and Pisoni (1986) employed a phonological priming paradigm in a lexical decision task. Although their primary results showed facilitation, the authors did note that in certain instances of high phonological similarity between primes and targets, some evidence of inhibition was observed. They speculated that these effects might arise from competition among phonologically similar lexical candidates.

Examples may be found in the literature of both facilitation and inhibition resulting from priming. It is of interest to consider what the fundamental differences are between studies like the present one and those such as Slowiaczek et al. (1987) that give rise to these differential effects. The only dimension distinguishing the studies is the level of priming—Slowiaczek et al. primed at a phonological level whereas we primed at a lower acoustic-phonetic level. Luce (1986) has suggested that the facilitation effects found by Slowiaczek et al. may have been due to expectancies generated by subjects during the course of the experiment. It is not unreasonable to make such an assertion; although the primes and targets in the Slowiaczek et al. study shared as little as one common phoneme, subjects may have easily noticed the consistent phonological relationships and generated their responses from a strategically restricted set of response alternatives.[3]

---

[3] If one considers these differential effects of priming in the context of an interactive activation model (e.g.

A number of studies have suggested that facilitation produced by priming arises from biases. These biases may not necessarily exert their influences at a conscious response level as active decisions, but may be present as perceptual biases. For example, Becker and Killion (1977) reported facilitation in a primed lexical decision task. Upon examining their results, the authors speculated that if subjects are induced to expect targets to occur from a small set of possibilites, they may be able to bypass a process of feature extraction, but only if the presented stimulus matches properties of the expected set. The stimuli for the present study were selected specifically to avoid this possible confounding. Our primes and targets shared no identical segments at all, making it very unlikely that subjects could learn any meaningful relations and strategically modify their guessing to take advantage of any cross-trial regularities.

The results of the present study strongly support the theoretical predictions of NAM. However, it is appropriate to examine alternate models of word recognition and to apply their predictions to the present data. As noted in Slowiaczek et al., several contemporary models of spoken word recognition cannot adequately account for effects of acoustic-phonetic priming. For instance, neither Forster's (1976) search theory [4] nor Klatt's LAFS model (1979) can explain the present results, or those of Slowiaczek et al. This is the case because these models include no mechanisms for comparing previously recognized words to new inputs; they only compare inputs to stored lexical representations in memory. Similarly, Morton's (1969) logogen theory is not equipped to account for the present results because logogens are not supposed to effect the thresholds of other logogens in the system.

In a recent chapter, Marslen-Wilson (1987) has proposed a modified version of cohort theory. Because the original version of the theory never directly addressed issues of word frequency, Marslen-Wilson has conducted several cross-modal semantic priming lexical decision experiments with words equated for recognition points. From his results, Marslen-Wilson argues that frequency effects arise early in the perceptual process. Specifically, he claims that high frequency words are *activated* relatively faster than low frequency words in the same cohort, so they are recognized faster. Additionally, he has recently added some com-

---

McClelland & Rumelhart, 1987) another explanation is suggested. These models predict that a prime will have the effect of encouraging recognition of items that share their particular features. If this were indeed the case, easily-recognized primes which share common phonemes with targets should increase the probability of target identification a priori, simply because the prime encourages guesses that share its features. It is not clear, therefore, whether facilitation from priming in an interactive activation model would represent a system in which activation from the prime actually facilitates activation of the target or if the experimental context of similar primes and targets simply correctly biases guessing.

[4] Hillinger (1980) has pointed out that Forster's (1976) search model predicts that a 'trail' of residual activation will be left after searching the peripheral access files. This feature of the model makes priming predictions possible. However, since Forster's access files are organized by frequency, one would expect that low frequency words could prime high frequency words, but not vice-versa. The present study obtained priming effects despite frequency ordering (low frequency primes affected both high and low frequency targets).

petitive processes to cohort theory which are similar to those proposed in NAM, stating that low frequency words are not only activated more slowly than high frequency words, but that they cannot be recognized until their high frequency competitors have dropped below some criterial level of activation. These latest assumptions make it very difficult, however, to determine whether or not cohort theory can now account for the differential results of prime frequency obtained in the present study. The implication that high frequency words receive immediate strong activation may lead one to predict that high frequency primes would tend to overshadow subsequent targets more than low frequency primes. Experiments 1A and 1B of the present study obtained exactly the opposite results. This disparity suggests that prime frequency may not be represented as a simple coding in the item's activation level. Low frequency primes seem to leave greater residual activation in the target's similarity neighborhood, and this extra activation translates into greater inhibition of target recognition performance. It may be necessary to employ sophisticated simulation analyses to adequately determine what cohort theory's predictions should be in this kind of situation.

In summary, the present study has demonstrated that priming a stimulus word with a phonetically related word will inhibit stimulus recognition. This effect is particularly robust when the prime items are low frequency words. The present study has also demonstrated that structural properties of lexical similarity neighborhoods are powerful predictors of spoken word recognition performance. These findings, taken together with the earlier findings of Luce and Pisoni (1988), provide additional support for the Neighborhood Activation Model. This model assumes that spoken words are recognized by initially activating a set of acoustically similar words in memory and then selecting from this set the item that is most consistent with the acoustic-phonetic information in the speech waveform. Word frequency is assumed to operate in the decision process to bias responses toward the more frequent lexical items in the activated neighborhood. We believe these attributes are primarily responsible for the speed, accuracy, and efficiency of the human word recognition system.

References

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Becker, C.A., & Killion, T.H. (1977). Interaction of visual and cognitive effects in word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, **3**, 389-401.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior*, **12**, 335-359.

Collins, A., & Loftus, E. (1975). A spreading activation theory of semantic processing. *Psychological Review*, **82**, 407-428.

Forster, K.I. (1976). Accessing the mental lexicon. In R.J. Wales & E. Walker (Eds.), *New approaches to language mechanisms*. Amsterdam: North Holland.

Hillinger, M.L. (1980). Priming effects with phonemically similar words: The encoding-bias hypothesis reconsidered. *Memory and Cognition*, **8**, 115-123.

Jakimik, J., Cole, R.A., & Rudnicky, A.I. (1985). Sound and spelling in spoken word recognition. *Journal of Memory and Language*, **24**, 165-178.

Klatt, D.H. (1979). Speech perception: a model of acoustic-phonetic analysis and lexical access. *Journal of Phonetics*, **79**, 279-312.

Kučera, F., & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6*. Bloomington, IN: Department of Psychology, Indiana University.

Luce, P.A., & Carrell, T.D. (1981). Creating and editing waveforms using WAVES. *Research on speech perception progress report no. 7*. Bloomington, IN: Department of Psychology, Indiana University.

Luce, P.A., & Pisoni, D.B. (1988). Neighborhoods of words in the mental lexicon. Manuscript under review.

Luce, R.D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71-102.

Marslen-Wilson, W.D., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

Marslen-Wilson, W.D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continous speech. *Cognitive Psychology*, **10**, 29-63.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McClelland, J.L., & Rumelhart, D.E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, **88**, 375-407.

Meyer, D.E., Schvaneveldt, R.W., & Ruddy, M.G. (1974). Functions of graphemic and phonemic codes in visual word recognition. *Memory and Cognition*, **2**, 309-321.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, **76**, 165-178.

Neely, J.H., Schmidt, S.R., & Roediger, H.L. III (1983). Inhibition from related primes in recognition memory. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **9**, 196-211.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10.* Bloomington, IN: Department of Psychology, Indiana University.

Slowiaczek, L.M., Nusbaum, H.C., & Pisoni, D.B. (1987). Phonological priming in auditory word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, **13**, 64-75.

Slowiaczek, L.M., & Pisoni, D.B. (1986). Effects of phonological similarity on priming in auditory lexical decision. *Memory and Cognition*, **14**, 230-237.

Tanenhaus, M.K., Flanigan, H.P., & Seidenberg, M.S. (1980). Orthographic and phonological activation in auditory and visual word recognition. *Memory and Cognition*, **8**, 513-520.

Taraban, R., & McClelland, J.L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language*, **26**, 608-631.

Webster's Seventh Collegiate Dictionary. (1967). Los Angeles: Library Reproduction Service.

Some Effects of Time-Varying Context on the Perception of Speech and Nonspeech Sounds [1]

John W. Mullennix, David B. Pisoni, and Stephen D. Goldinger

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported by NIH Research Grant *NS-12179-12* and NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN. The authors wish to thank Keith Johnson, Chris Martin, and Van Summers for comments on earlier versions of the manuscript and Amy Lawlor for assistance in running subjects.

Abstract

In a recent investigation, Williams (1986) obtained evidence inconsistent with an auditory-based processing account of consonantal context effects in vowel perception. In the present experiments, the perception of vowels in a dynamically-changing context was examined further. Two 20-item synthetic vowel series ranging from [U]-[I] and [wUw]-[wIw] were generated along with three-component sinewave nonspeech continua based on the two speech series. AXB categorization was obtained under four experimental conditions: Speech vowel, sinewave vowel, speech pitch, and sinewave pitch. The results showed that categorization shifts occurred between the steady-state and time-varying continua within each stimulus condition, demonstrating the influence of the immediately surrounding acoustic context. In a second experiment, when the AXB endpoints were crossed, categorization shifts opposite in direction to those observed in the first experiment were obtained. Both sets of findings failed to replicate Williams' earlier results and call into question his interpretation in terms of a uniquely phonetic mode of processing. The results are interpreted as additional support for an auditory-based account of context effects in the perception of both speech and nonspeech signals.

Some Effects of Time-Varying Context on the Perception of Speech and Nonspeech Sounds

Over the last few years, the perception of speech has been approached from several different theoretical perspectives (Cole & Scott, 1974; Elman & McClelland, 1986; Fowler, 1986; Liberman, Cooper, Shankweiler, & Studdert-Kennedy, 1967; Liberman & Mattingly, 1985; Oden & Massaro, 1978; Pastore, 1981; Pisoni & Luce, 1987; Schouten, 1980). According to one account, speech perception involves intimate links with speech production (Liberman, 1970a, 1982; Liberman et al., 1967; Liberman & Mattingly, 1985; Studdert-Kennedy, Liberman, Harris, & Cooper, 1970). By this view, the listener has perceptual mechanisms that have access to knowledge of intended articulatory "gestures" specific to speech utterances and this articulatory knowledge mediates perception (Liberman & Mattingly, 1985). This view may be contrasted with another approach that assumes that speech perception involves processes that are used for the perception of both speech and nonspeech sounds through the use of general auditory processing mechanisms (Cutting, 1978; Massaro, 1987; Pastore, 1981; Schouten, 1980). According to this view, articulatory knowledge has little or no role in the perception of speech. Instead, the perceptual processes simply encode auditory parameters or features contained in the incoming speech signal and map these features onto phonetic categories stored in long-term memory.

Although there are other approaches to speech perception that do not fall into this simplified dichotomy (e.g., Elman & McClelland, 1986; McClelland & Elman, 1986), much of the experimental research conducted in speech perception over the last 20 years has been directed towards testing the claims inherent in articulatory-based and auditory-based accounts of speech perception. A great deal of research devoted to specific perceptual phenomena such as categorical perception (e.g., Liberman, Harris, Hoffman, & Griffith, 1957; Mattingly, Liberman, Syrdal, & Halwes, 1971), cerebral lateralization (e.g. Studdert-Kennedy & Shankweiler, 1970), auditory-visual integration (Massaro, 1987; McDonald & McGurk, 1978; McGurk & McDonald, 1976; Summerfield, 1979), phonetic trading relations (see Repp, 1982; 1983), and duplex perception (Liberman, Isenberg, & Rakerd, 1981; Repp, 1984; Whalen & Liberman, 1987) has been often cited as support for some type of articulatory-based or motor-based account of speech perception. Proponents of this view argue that these perceptual phenomena arise from the operation of specialized, speech-specific perceptual processes related to articulatory knowledge or constraints (see also Liberman & Mattingly, 1989).

Support for auditory-based theories of speech perception has largely rested on experimental evidence obtained using nonspeech control stimuli. Typically, these stimuli are designed to model the acoustic structure of phonetic distinctions in natural speech such as voice onset time, formant frequency trajectories, or variations in rate of change of formant transitions (e.g., Miller, Wier, Pastore, Kelley, & Dooling, 1976; Pisoni, 1977; Pisoni, Carrell, & Gans, 1983). The similarity of the results obtained in studies using nonspeech control stimuli to the results of studies using speech stimuli has been offered as evidence against the postulation of specialized perceptual mechanisms in speech perception (Hillenbrand, 1984; Miller et al., 1976; Parker, 1988; Parker, Kleunder, & Diehl, 1986; Pastore, Harris, & Kaplan, 1982; Pisoni, 1977; Pisoni et al., 1983; Ralston, 1986; Ralston & Sawusch, 1984). For the most part, arguments for auditory-based theories are based on the logic that the perceptual phenomena obtained using nonspeech auditory signals are due to processes that are not speech-specific but are instead related to the general auditory processing capabilities of the listener.

Despite the large number of studies in speech perception that have been conducted with these two opposing theoretical viewpoints in mind, the question of whether speech perception may be mediated by mechanisms related to articulatory knowledge or by mechanisms based on general auditory processes has not yet been satisfactorily resolved. In order to examine these issues further, it is worthwhile to extend further investigations to encompass the perception of vowels occurring in speech, as most of the previous research has concentrated on perception of consonants (see above). In particular, if vowel perception is examined under conditions where the effects of context due to adjacent-occurring consonants can be observed, this provides an opportunity to test the predictions of a particular framework of speech perception in a situation that preserves some important contextual components of fluent speech.

It is well-known that the acoustic cues to phonetic distinctions are not represented in the waveform in a straightforward manner (Liberman, Delattre, & Cooper, 1952; Liberman, Delattre, Cooper, & Gerstman, 1954). Coarticulatory influences on the production of consonants and vowels result in a complex mapping of acoustic cues to phonetic segments (Liberman, 1970b; Liberman et al., 1967). There are numerous well-known examples of perceptual phenomena in speech that illustrate the complexity of the mapping of acoustic cues to phonetic segments (e.g., Liberman et al., 1952; 1954). In the present study, we chose to focus on one particular aspect of recovering segmental information from the speech signal. That is, we examined the extraction of vowel information when vowel cues are embedded in dynamically changing contexts. By examining this problem in detail, we hoped to provide some experimental evidence to help determine the appropriateness of either an articulatory-based or an auditory-based model of speech perception.

Vowels that are produced by speakers of English have a number of well-defined acoustical characteristics. When vowels are produced in isolation, they contain steady-state formant frequency values. But when vowels are produced in context with other adjacent phonetic segments, coarticulatory influences on the vowel often result in formant frequencies that deviate substantially from their steady-state values (Lindblom, 1963; Stevens & House, 1963). In particular, when a vowel is produced in the context of two surrounding consonants, the resulting formant frequencies are typically "reduced." This phenomenon is often called "articulatory undershoot." The vowel region is not only reduced in length but contains formant frequency values that are either higher or lower than those obtained when the vowel is produced in isolation. However, even though the acoustic cues for the vowel change when it is produced in dynamic context, the perceptual system "recognizes" the vowel as the same phonetic segment regardless of these acoustic differences. Given this phenomenon, it is of interest to determine the nature of the perceptual mechanisms that are responsible for recovering the intended vowels. There is some evidence from a series of studies conducted by Strange and colleagues (see Jenkins, Strange, & Edman, 1983; Strange, 1987; Strange, Jenkins, & Johnson, 1983) indicating that coarticulated vowels are identified quite accurately even when the steady-state region of the vowel is silenced. That is, vowels are identified by virtue of only the formant transitions into and out of the intended vowel. Their results provide support for the hypothesis that the acoustic information contained in the surrounding formant transitions is used by the perceptual system in perceiving particular vowels.

The effects of context on vowel perception were investigated more than 20 years ago in a well-known study conducted by Lindblom and Studdert-Kennedy (1967). Lindblom and Studdert-Kennedy created three synthetic vowel continua, one consisting of steady-state isolated vowels ranging from [U]-[I], one consisting of [U]-[I] vowels embedded in a [wVw] consonantal context, and one consisting of [U]-[I] vowels embedded in a [jVj] consonantal context. Typically, in fluent coarticulated speech, vowel formant frequencies are lowered by a labial [wVw] context and raised by a palatal [jVj] context. In Lindblom and Studdert-Kennedy's study, however, the target vowel formant frequency values across all series were identical. The authors hypothesized that if the immediately surrounding acoustic context provided by the consonantal formant transitions adjacent to the vowel affected perception, then the categorization of the vowel, as reflected by the location of the category boundaries for the [wVw] series and for the [jVj] series, should be shifted relative to the steady-state [U]-[I] series. This result would be anticipated if the acoustic cues contained in the formant transitions induce the perceptual system to "compensate" for vowel formant values that differ from those characteristic of the vowel produced in isolation. However, if context did not have this effect, the authors expected that the vowels from all three series would be categorized in exactly the same manner. The results obtained by Lindblom and Studdert-Kennedy showed that the category boundary for the [wVw] series was shifted to the left of the steady-state [U]-[I] series (that is, less "U" responses were obtained), while the category boundary for the [jVj] series was shifted to the right of the steady-state [U]-[I] series (more "U" responses were obtained). These results support the hypothesis that the perceptual system performs some type of perceptual compensation or re-adjustment based on the dynamic information contained in the consonantal formant transitions into and out of the target vowel. Apparently, in response to formant transition information that is generally accompanied by "articulatory undershoot", the perceptual system engages a perceptual mechanism that extrapolates beyond the actual vowel formant values in order to recover the intended vowel target ("perceptual overshoot").

How do the results obtained over 20 years ago by Lindblom and Studdert-Kennedy relate to the issue of articulatory-based versus auditory-based models of speech perception? Recall that articulatory-based theories assume that the perceptual system recovers phonemes by virtue of an invariant correspondence between the acoustic cues present in the signal and some representation of articulatory configurations related to producing the acoustic patterns. Recall also that auditory-based theories maintain that segmental units are recovered by auditory processes that extract the relevant phonetic information without recourse to articulatory knowledge. Lindblom and Studdert-Kennedy were not clear about precisely how their results could be interpreted in terms of articulatory-based versus auditory-based models of speech perception. They suggested that their results were compatible with "active" perceptual models such as analysis by synthesis (Stevens & Halle, 1967) and motor theory (Liberman et al., 1967), but they also entertained the idea that adaptation or contrast effects due to peripheral auditory processes could also account for their results. The results they obtained did not provide strong evidence supporting either view, thus, the precise nature of the compensation mechanism remained unclear at that time.

A recent series of categorization experiments has been carried out by Williams (1986) to provide an explicit test of the auditory processing account suggested by Lindblom and Studdert-Kennedy. Williams (1986) created a set of synthetic speech stimuli similar to the [U]-[I] steady-state vowel series and [wUw]-[wIw] vowel series of Lindblom and Studdert-Kennedy. He also generated two sets of sinewave tone-analog stimuli based on the frequency values of the stimuli contained in the [U]-[I] and [wUw]-[wIw] vowel series. The use of sinewave control stimuli preserves the dynamic time-varying information characteristic of speech but eliminates the acoustic cues to source characteristics (see Cutting, 1974; Remez, Rubin, Pisoni, & Carrell, 1981). Under certain experimental conditions, sinewave stimuli of this type can be heard either as nonspeech tones or as speech (e.g., Bailey, Summerfield, & Dorman, 1977; Best, Morrongiello, & Robson, 1981; Grunke & Pisoni, 1982; Ralston, 1986; Remez et al., 1981; Schwab, 1982; Tomiak, Mullennix, & Sawusch, 1987).

Williams (1986) used these sets of stimuli in three different experimental conditions. In one condition, listeners received synthetic steady-state [U]-[I] vowels and time-varying [wUw]-[wIw] speech stimuli and categorized them into the vowel categories [U] and [I]. In a second condition, listeners received steady-state and time-varying sinewave tone-analog stimuli. In this condition, they were told that the stimuli were speech sounds and they were to categorize them into the same two vowel categories as used in the speech condition. In the third condition, listeners were presented with the two types of sinewave stimuli but were told that the stimuli were nonspeech tones and they were to categorize the stimuli on the basis of pitch. Williams hypothesized that if this "perceptual overshoot" effect was due to some type of peripheral auditory constraint involved in processing the frequency spectrum over time, such as adaptation or contrast, then category boundary shifts should be observed in both the speech and nonspeech conditions for the sinewave control stimuli. However, if speech-specific coding processes rather than peripheral auditory processes were responsible for these effects, then boundary shifts should only be observed for the speech stimuli and for the sinewave stimuli heard as speech but not for the sinewave stimuli heard as nonspeech sounds.

Williams (1986) obtained category boundary shifts for the speech stimuli and for the sinewave stimuli heard as speech that were similar to the shifts obtained by Lindblom and Studdert-Kennedy. However, no category shift was obtained for the sinewave stimuli heard as nonspeech in the pitch categorization task. Williams interpreted these results as support against the claim that auditory processing constraints are responsible for the "perceptual overshoot" phenomenon found in the perception of coarticulated vowels. Instead, he suggested that the perceptual processing of coarticulated vowels must be at least partially based on articulatory information which is "encoded" into the speech signal.

On first inspection, the results of Williams (1986) appear to be consistent with the proposal that the processes involved in recovering phonetic segments from a dynamically time-varying speech signal are closely related to the use of articulatory information in speech perception. However, the major support for this claim relies on a null result. That is, support for Williams' hypothesis rests on the *failure* to find a boundary shift in the sinewave pitch condition. There are numerous reasons why an experimental effect may not be obtained, including methodological peculiarities, uncontrolled individual differences between subjects, experimental differences at the time of testing, etc. Although Williams made an explicit prediction a priori anticipating this null result, his assertions should be viewed with some caution pending replication.

Our motivation for conducting the first experiment in the present report was to further investigate the role of consonantal context in vowel perception. More specifically, the purpose was twofold. Our first goal was to replicate the results obtained by Lindblom and Studdert-Kennedy (1967) using speech stimuli and the more recent results obtained by Williams (1986) using speech stimuli and sinewave nonspeech tone-analogs. If we obtain a pattern of results identical to that obtained by Williams, then one could place more confidence in the validity of his basic results and focus discussion on the theoretical claims and implications. A replication would give additional support to the claim that the null result in the sinewave pitch condition actually reflected a difference in processing between speech and nonspeech conditions. However, if we fail to replicate Williams' results using identical testing procedures and identical stimuli, this finding would raise questions about Williams' experimental data and his subsequent interpretations.

Our second goal was to investigate the effects of selective attention to particular attributes of the stimulus on AXB categorization. In Williams' sinewave pitch condition, listeners were specifically told to attend to the pitch of the stimuli in making their responses. For the sinewave vowel and speech vowel conditions, on the other hand, listeners were instructed to attend to the vowel quality of the stimulus and to make their responses with regard to the identity of the vowel. Given these instructions, it is possible that the differences found between the nonspeech sinewave pitch condition and the speech conditions were simply due to differences in selective attention to different stimulus dimensions and not to some fundamental difference in mode of processing. For the speech conditions, although the identity of the vowel is correlated with differences in formant frequency values, it is not clear what other aspects of the speech signal are actually used in order to make a categorization judgment. For the sinewave pitch condition, however, only frequency could be used in making a pitch categorization response. Thus, the differences Williams found between speech and nonspeech may have been due to differences in selective attention to either vowel quality or pitch and not to the employment of fundamentally different perceptual mechanisms.

In order to examine the role that selective attention to specific stimulus aspects may play in categorization, we added an additional control condition. In this condition, called the speech pitch condition, listeners were given speech stimuli to categorize. However, instead of focusing the subjects' attention on vowel color, subjects were specifically instructed to attend to the pitch of the vowel and to make their categorization responses on the basis of pitch alone. Thus, the listener's attention was specifically directed to the pitch quality of the vowel. This additional condition provides a direct assessment of whether selective attention to different aspects of the vowel has any effect on perception. In addition, a comparison of performance between the speech pitch condition and the sinewave pitch condition provides an opportunity to directly assess any effects due to speech/nonspeech stimulus differences when pitch is attended to.

Several outcomes are possible. If a categorization shift is obtained for both vowel conditions and the speech pitch condition, this finding would suggest that the results obtained by Williams were not due to differences in task demands but actually reflected a genuine difference in processing mode for speech and nonspeech sounds. On the other hand, if a categorization shift is obtained for the two vowel conditions but is absent for the two pitch conditions, this would suggest that the results observed by Williams may have been due to the nature of the information attended to by listeners in performing the categorization task. A third possible outcome is that a category boundary shift is obtained in all four experimental conditions. Such a pattern of results would suggest that the null result in the sinewave pitch condition obtained by Williams may have been due to chance and that an effect of context was actually present but was obscured by some factor or factors. This finding would also suggest that the processes which produce "perceptual overshoot" in the perception of coarticulated vowels operate for both speech and nonspeech sounds regardless of how attention is allocated to certain aspects of the signal.

Experiment 1

Method

Subjects. Sixty-seven undergraduate students from introductory psychology courses at Indiana University served as subjects. Each subject took part in one 1-hour session and received partial course credit for participating in the experiment. All subjects were native speakers of English who reported no history of a speech or hearing disorder at the time of testing.

Stimulus Materials. The stimuli consisted of synthetic steady-state vowels, synthetic consonant-vowel-consonant (CVC) nonsense syllables, and three-component sinewave tone-analogs of speech based on the stimuli used by Williams (1986). The speech stimuli were generated with the software synthesis program developed by Klatt (1980). A steady-state vowel continuum was created by varying the formant frequencies in 20 steps from the endpoint [U] to the endpoint [I]. Each stimulus was 100 msec in duration. The [U] vowel endpoint contained steady-state formant frequencies of 350 Hz for F1, 1000 Hz for F2, and 2300 Hz for F3. The [I] vowel endpoint contained formant values of 350 Hz for F1, 2000 Hz for F2, and 2825 Hz for F3. The 20-item series was created by incrementally changing the values of F2 and F3 through 19 steps by the use of an algorithm adopted from Lindblom and Studdert-Kennedy (1967). The values for F1 remained fixed at 350 Hz for all stimuli. The algorithm is given below:

```
F1 = 350
F2 = 1000 + [(n-1) x 1000] / 19
F3 = 2300 + [(n-1) x  525] / 19

(n = stimulus number)
```

The CVC nonsense syllable continuum was created by taking the formant frequency values for each stimulus in the steady-state vowel continuum and using them as target frequency values for the corresponding vowel in the CVC continuum. For each CVC stimulus, a parabolic formant frequency trajectory was computed beginning at and ending at a set of frequency loci on each side of the vowel appropriate for the semivowel [w]. The loci values were 250 Hz for F1, 800 Hz for F2, and 2200 Hz for F3. The [wVw] stimuli contained continuously varying formants which began at one locus, increased up to the vowel target frequency values at the midpoint of the stimulus, and then decreased back down to terminate at the other locus. The formant trajectories for each stimulus were parabolic in shape and symmetrical on each side of the vowel midpoint region. A 20-item continuum was created ranging perceptually from [wUw] to [wIw]. Each stimulus in the continuum contained vowel target formant frequency values identical to the values for the corresponding stimulus in the steady-state vowel continuum. Each [wVw] stimulus was 100 msec in duration.

The formant frequencies for F4 and F5 were held at 3500 Hz and 4500 Hz for all stimuli. Bandwidths were set at 90 Hz for F1 and F2, 180 Hz for F3, and 300 Hz for F4 and F5. Fundamental frequency was held at a constant 90 Hz and voicing amplitude was specified at a constant 60 dB throughout the duration of the stimuli.

Two sets of 20 nonspeech sinewave tone-analog stimuli were also created. One set was based on the steady-state [U]-[I] vowel continuum and one set was based on the [wUw]-[wIw] continuum. The tone-analog stimuli were generated with a program that synthesizes three-component sinewave tones (Kewley-Port, 1976). The frequency values specified for F1, F2, and F3 in the speech stimuli of both continua were used as the frequency values for each set of the nonspeech tone-analog stimuli. In order to specify relative amplitude values for the three tone components corresponding to the relative amplitude values between formants in the speech stimuli, the amplitudes of F1, F2, and F3 for each of the synthetic speech stimuli were determined using spectral analysis techniques. The relative amplitudes obtained for the formants for each stimulus were used to specify the amplitude values for the tone components in the corresponding tone-analog stimuli. After generating all of the stimuli, RMS amplitude levels for the speech and tone stimuli were digitally equated using a software package designed to modify speech waveforms.

Procedure. Four experimental conditions were formed: Speech vowel (SV), sinewave vowel (SWV), speech pitch (SP), and sinewave pitch (SWP). The experimental procedure used an AXB identification task. In the AXB task, listeners are presented with sequences of three stimuli and are asked to indicate whether the second stimulus "X" sounds more like the "A" stimulus or more like the "B" stimulus in the sequence. The "A" and "B" stimuli are the two endpoint stimuli of the continuum being tested. In the present experiment, all 20 members of a particular stimulus continuum were presented as "X" stimuli. Subjects were instructed to push a button on a response box labeled "first" if they thought the "X" stimulus sounded more like the first stimulus, and to push a button labeled "third" if they thought "X" sounded more like the third stimulus. In this manner, identification categorization responses were obtained for each stimulus item in each continuum without the use of explicit labels.

Within each experimental condition, listeners were presented with stimuli from a steady-state continuum in one set of trials and with stimuli from a time-varying continuum in a separate set of trials. In the SV condition, listeners were presented with the [U]-[I] and [wUw]-[wIw] speech stimuli. Subjects in this condition were told that they would be listening to high-quality synthetic speech sounds generated by a computer. For the steady-state continuum, they were instructed to make their responses on the basis of whether the "X" vowel stimulus sounded more like the "A" vowel endpoint or the "B" vowel endpoint. For the time-varying continuum, they were also told to make their responses on the basis of whether the vowel contained in the "X" stimulus sounded more like the "A" vowel endpoint syllable or the "B" vowel endpoint syllable.

In the SWV condition, listeners were presented with steady-state and time-varying sinewave tone-analogs based on the [U]-[I] and [wUw]-[wIw] speech continua. However, for both sets of stimuli, listeners were told that they would be listening to synthetic speech sounds generated on a computer. Furthermore, they were told that even though the stimuli may sound somewhat unnatural and unlike normal human speech, they were indeed listening to speech sounds consisting of isolated vowels and nonsense syllables. They were instructed to make their responses on the basis of whether the "X" stimulus vowel sounded more like the "A" stimulus vowel or more like the "B" stimulus vowel, just as the instructions were specified for listeners in the SV condition.

In the SP condition, listeners were also presented with the [U]-[I] and [wUw]-[wIw] speech stimuli. However, they were instructed to listen carefully to the pitch of the vowel and to make their responses on the pitch quality of the vowel. For example, for the steady-state continuum, if the pitch of the "X" vowel stimulus sounded closer in pitch to the [U] vowel endpoint, they were to respond by pushing the button appropriate for the [U] endpoint. Likewise, they were told to use the same criteria in responding to the [wUw]-[wIw] stimuli.

Finally, in the SWP condition, listeners were presented with the sinewave tone-analog continua. For both the steady-state and time-varying stimuli, listeners were told that they would hear sounds which were composed of tones. They were instructed to listen to the pitch of the sounds and to make their AXB responses on the basis of the pitch of each stimulus.

For each experimental session, three blocks of trials were presented for each set of stimuli. The first set of three blocks consisted of stimuli from one continuum and the second set consisted of three blocks of stimuli from the other continuum. The order of the "A" and "B" endpoints across trials within each set was randomized and presentation of the "X" stimuli was randomized. The order of continua within each session was counterbalanced across subjects. Half of the subjects received stimuli from the steady-state continuum in the first three blocks and one half received stimuli from the time-varying continuum in the first three blocks. The first block in each set consisted of 20 AXB trials with one repetition of each of the 20 stimulus items in the continuum as the "X" stimuli. Listeners were told that these trials were practice, and they were not included in the data analyses. The second two blocks in each set consisted of 100 AXB trials each, produced by five repetitions of the 20 stimuli occurring as "X". Thus, a total of 10 identification responses per stimulus was collected from each subject.

A warning light on each response box was illuminated before the beginning of each trial sequence. A 1-sec interval elapsed between offset of the warning light and presentation of the first stimulus. Each stimulus in the sequence was separated by a 500 msec inter-stimulus-interval. Each new trial was initiated either after all subjects had entered a response or after a 3-sec interval had elapsed. No feedback was given to the subjects. There was a 60-sec interval between each block of trials. Subjects were run in small groups ranging from 3-4 subjects per group. Stimulus output and data collection were controlled on-line by a PDP-11/34A computer. Stimuli were output via a 12-bit digital-to-analog converter at a 10 kHz sampling rate and were low-pass filtered at 4.8 kHz before presentation to subjects via matched and calibrated TDH-39 headphones.

Results

At the end of each experimental session, subjects were asked to complete a post-test questionnaire (see Appendix A). The questionnaire was designed to obtain information about how the stimuli sounded to the subject and what types of response strategies subjects employed in making their responses. The questionnaire included questions about what the stimuli sounded like, whether they sounded like speech or not, what kind of strategies or stimulus properties the subject used in order to make responses, etc. For the SV and SWV conditions, if a subject indicated on any question on the questionnaire that the stimuli did not sound speechlike or that the vowel class of the stimuli was not used in order to make responses, the subject's data were eliminated from further data analysis. For the SP condition, if a subject indicated on any question that the stimuli did not sound like speech or that the pitch of the vowel was not used to make responses, the subject was eliminated. For the SWP condition, if a subject indicated on any question that the stimuli sounded like speech or that the pitch of the stimulus was not used as a basis to make responses, the subject was eliminated. As a result of these stringent selection criteria, the data from 11 out of 12 subjects were analyzed for the SV condition, 11 out of 25 subjects for the SWV condition, 10 out of 15 subjects for the SP condition, and 11 out of 15 subjects for the SWP condition.

Insert Table 1 and Figure 1 about here

For each subject, the data were tabulated in terms of the percentage of times each stimulus was judged to be like the "stimulus 1" endpoint. The best-fitting normal ogive through these points was then determined (Woodworth, 1938). The 50% point of this ogive was taken as the category boundary crossover point in the categorization function. The 50% crossover points were used as dependent measures entered into the statistical analyses conducted on the data (see Table 1). Figure 1 shows the overall mean percent "stimulus 1" responses collapsed over subjects for the steady-state [U]-[I] and the time-varying [wUw]-[wIw] continua in the SV, SWV, SP, and SWP conditions. As shown in the figure, the categorization

Table 1

Individual subjects' category crossover point values for the steady-state (SS) and time-varying (TV) continua for all conditions in Experiment 1. The overall mean crossover values and standard deviations (slope values) for each condition are also displayed.

	Speech Vowel		Sinewave Vowel		Speech Pitch		Sinewave Pitch	
Ss	SS	TV	SS	TV	SS	TV	SS	TV
1	11.7	9.2	10.4	9.1	9.8	9.0	9.6	8.2
2	10.0	9.4	10.5	7.3	12.5	11.4	11.6	8.9
3	10.7	10.3	9.2	8.5	11.4	9.8	10.7	8.0
4	12.5	9.8	11.2	11.4	11.6	10.2	10.7	10.2
5	10.2	9.5	10.1	8.9	13.5	12.4	10.6	9.0
6	10.9	9.2	11.1	10.0	12.6	11.0	10.8	11.8
7	12.1	6.9	9.8	9.3	10.6	9.6	11.6	9.7
8	10.2	9.2	9.6	9.6	10.1	10.1	10.0	10.3
9	10.7	6.9	10.4	11.3	11.4	9.0	10.0	8.6
10	11.3	10.5	9.0	10.1	11.0	9.4	9.5	8.3
11	10.9	8.6	9.5	8.7	****	****	10.9	8.9
Mean	11.0	9.0	10.1	9.5	11.4	10.2	10.6	9.3
S.D.	-4.0	-3.8	-3.7	-3.6	-3.8	-3.8	-4.0	-3.7

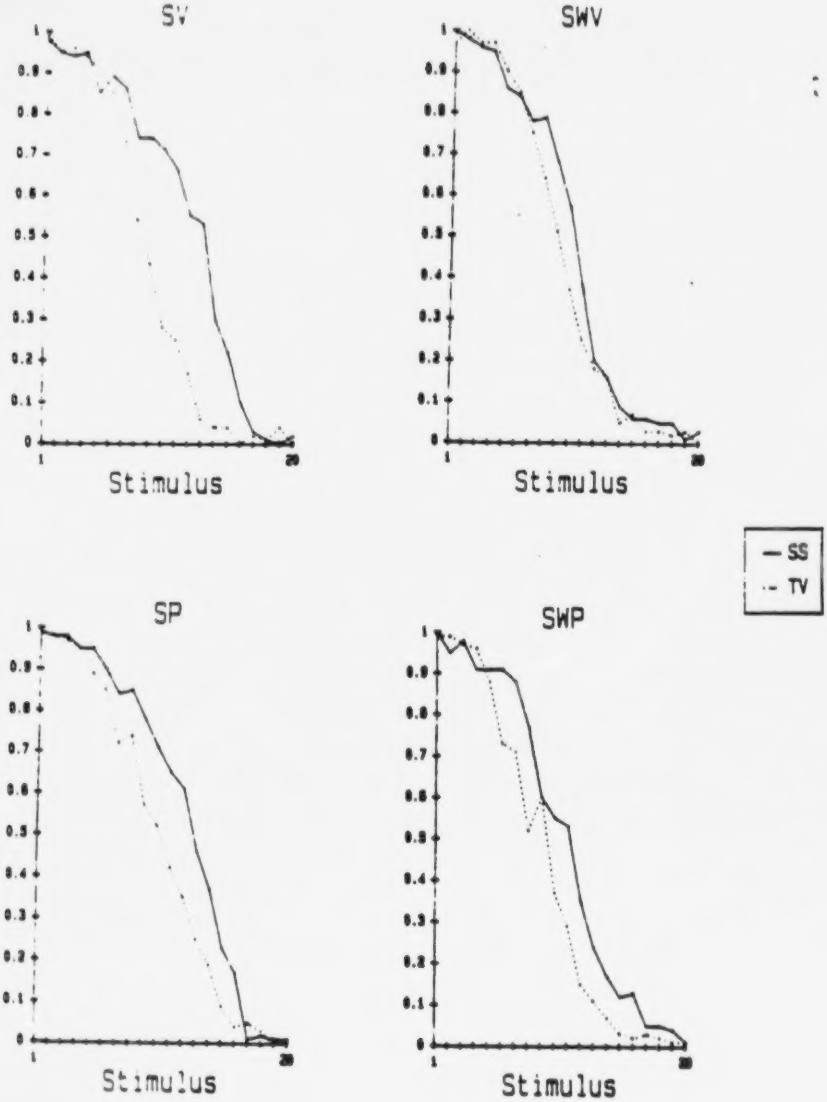


Figure 1. AXB categorization functions averaged over subjects for the steady-state [U]-[I] and context-dependent [wUw]-[wIw] continua for the SV, SWV, SP, and SWP conditions for Experiment 1.

function for the time-varying [wUw]-[wIw] continuum is shifted to the left of the steady-state [U]-[I] continuum in all four experimental conditions. Separate analyses were carried out to assess the effect of context in each of the four conditions. These are reported below.

Speech Vowel

A one-way ANOVA was conducted on the data for the SV condition. A significant main effect of continuum was obtained $\underline{F}(1,10) = 18.5$, $\underline{p} < .01$. The category boundary crossover value for the [wUw]-[wIw] continuum was significantly less than the crossover value obtained for the steady-state [U]-[I] continuum (9.0 and 11.0, respectively). The shift in the AXB categorization function observed as a result of the [wVw] consonantal context is similar to shifts in categorization reported previously by Lindblom and Studdert-Kennedy (1967) and Williams (1986).

Sinewave Vowel

A one-way ANOVA was also conducted on the data for the SWV condition. A significant main effect of continuum was not obtained $\underline{F}(1,10) = 3.0$, $\underline{p} < .12$. However, the category boundary crossover value for the [wUw]-[wIw] tone-analog continuum was again less than the crossover value for the [U]-[I] tone-analog continuum (9.5 and 10.1, respectively). The context provided by the tone-analog components corresponding to [w] resulted in a categorization shift similar to the shift obtained in the SV condition, however, the difference was not statistically significant.

Speech Pitch

A one-way ANOVA was also conducted on the data for the SP condition. A significant main effect of continuum was observed $\underline{F}(1,9) = 39.8$ $\underline{p} < .001$. The category boundary crossover value for the [wUw]-[wIw] continuum was significantly less than crossover value for the [U]-[I] continuum (10.2 and 11.4, respectively). Thus, as shown by the AXB categorization shift, when subjects made their responses on the basis of the pitch of the vowel, the consonantal context provided by the [w] semivowel produced a substantial and significant effect on categorization performance.

Sinewave Pitch

Finally, a one-way ANOVA was conducted on the data for the SWP condition. A significant main effect of continuum was obtained $\underline{F}(1,10) = 13.9$, $\underline{p} < .01$. The category boundary crossover value for the [wUw]-[wIw] tone-analog continuum was significantly less than the crossover value for the [U]-[I] tone-analog continuum (9.3 and 10.6, respectively). Thus, the context provided by the surrounding [w] tone-analog transitions also affected AXB categorization in the nonspeech pitch task.

Discussion

The results obtained in Experiment 1 provide further evidence concerning the processes involved in the perception of coarticulated vowels in CVC syllables. First, we successfully replicated the findings of Lindblom and Studdert-Kennedy (1967). Significant categorization shifts were obtained between isolated, steady-state vowels and vowels embedded in symmetrical consonantal contexts. This result demonstrates that the perceptual system uses dynamic time-varying acoustic information contained in the formant transitions surrounding vowels to support recovery of the cues needed for recognition of the target vowels.

Second, we obtained some results that differed from those obtained by Williams (1986). Unlike Williams, we also observed a categorization shift in the sinewave pitch condition, in which listeners categorized the pitch of the nonspeech stimuli. This result contrasts with Williams' failure to find a shift in his sinewave pitch condition. Given that our stimuli and categorization task were identical to those used by Williams, it is not immediately obvious to us why we obtained this finding. However, this result is consistent with the idea that the null result obtained by Williams in his sinewave pitch condition belied what was actually a perceptual effect of context.

We also obtained a significant categorization shift in the speech pitch condition, in which listeners judged the pitch of the vowel. This shift was similar to the shifts obtained in the other conditions. This result suggests that context affects vowel categorization processes even when attention is directed towards the pitch attributes of vowels instead of vowel color. Thus, at least within the present experiment, focusing subjects' attention to particular attributes of the speech stimulus did not appear to have a differential impact on the perceptual processes involved in extracting vowel target information from context.

Finally, although we obtained significant categorization shifts between the steady-state isolated vowel stimuli and the time-varying vowel stimuli using speech stimuli, we did not obtain a significant categorization shift when using sinewave tone-analogs heard as speech. However, the direction of shift in the categorization function for the SWV was similar to that of the other conditions. One could speculate that the reason for this result is that the richer set of acoustic cues present in natural speech may contribute in some unspecified manner to the magnitude of the context effect. Thus, when using nonspeech stimuli heard as speech, the effect may not be as salient.

Experiment 2

Although the results obtained in Experiment 1 demonstrated that the time-varying context present in speech and nonspeech signals has similar effects on perception, we did not directly assess the manner in which the information contained in time-varying stimuli was processed. Williams (1986) noted that for the sinewave pitch judgments, two different processes may have produced his results. He suggested that listeners in the sinewave pitch condition could have perceptually isolated the maximum target frequencies of the time-varying stimuli in order to extract pitch. Alternatively, they may have integrated the frequency information across the durations of the stimuli in order to extract pitch. In order to distinguish between these two alternatives, Williams performed an additional AXB categorization experiment in which the "A" and "B" endpoints for each continuum were "crossed." In other words, instead of the time-varying stimuli and steady-state stimuli being judged against endpoint stimuli drawn from their own continuum, the time-varying stimuli were judged against the "A" and "B" endpoints from the steady-state continuum and the steady-state stimuli were judged against the "A" and "B" endpoints from the time-varying continuum. Williams hypothesized that if listeners were integrating frequency information over the stimulus, the time-varying stimuli should be judged as lower in pitch when judged against the endpoints from the steady-state continuum, and the steady-state stimuli should be judged as higher in pitch when they were judged against endpoints from the time-varying continuum. The results Williams obtained for the sinewave pitch condition showed that the category boundary for the time-varying continuum was shifted to the right of the steady-state continuum, a shift that was opposite in direction to the shifts observed in his previous experiments. This result was consistent with the proposal that listeners integrated the frequency information over the stimulus in order to extract pitch information. However, the cross-series manipulation did not affect categorization in the two vowel conditions, as significant leftward shifts of the time-varying continuum boundary were obtained just as before. Thus, the category shift obtained in the sinewave pitch condition was opposite in direction to the shifts obtained in the vowel conditions.

The difference in categorization performance between speech and nonspeech conditions observed as a function of the cross-series manipulation was taken by Williams as evidence that two different types of perceptual processes produced the results. Williams suggested that listeners' pitch judgments in the nonspeech condition were based on the "unitary" or "static" aspects of the signal that are obtained by integrating frequency over time. According to Williams, differences between speech and nonspeech conditions arose because vowels are perceived by different processes than those used in pitch perception. Since an integrative-type process accounted for the nonspeech results, and since the pattern of results differed between speech and nonspeech, Williams concluded that the perception of coarticulated vowels must be based on processes that encode dynamic time-varying information in the speech signal.

The rationale offered for a dynamic explanation of coarticulated vowel perception must be carefully considered, however. The sole support for this particular assertion rests simply on the fact that differences in patterns of performance existed between speech and nonspeech conditions. At best, these results provide only indirect evidence for the operation of dynamic processes in speech perception. Direct evidence indicating that the processes involved in the perception of coarticulated vowels are dynamic was not obtained; the presence of dynamic processes was only *inferred* from the pattern of speech/nonspeech results. However, based on the importance of the performance differences exhibited as a result of the cross-series manipulations, we decided to replicate these findings in order to assess their robustness. In addition, we included an additional experimental condition in order to further elucidate the nature of the processes involved in perception of vowels embedded in dynamic context.

In this experiment, we adopted the cross-series methodology used by Williams (1986). We created four experimental groups: Speech vowel crossed (SV-crossed), sinewave vowel crossed (SWV-crossed), speech pitch crossed (SP-crossed), and sinewave pitch crossed (SWP-crossed). The changes in experimental design produced by using the cross-series manipulation were similar to those of Williams, with the "A" and "B" endpoints for each AXB task drawn from the continuum opposite that of the stimuli used as "X" test items. In addition to the three experimental conditions tested by Williams, we included an additional condition called the speech pitch-crossed condition (SP-crossed), in which listeners judged the pitch of the vowels. As in Experiment 1, the inclusion of the SP-crossed condition in the present experiment provided an important control condition.

The possible outcomes of Experiment 2 and their interpretations are as follows. If categorization shifts are obtained that are similar to those observed in Williams' experiment, then the reliability of his results will be given further support. If these results are obtained, then the results of the SP-crossed condition become important. If the categorization shift in the SP-crossed condition is similar to the shifts obtained in Williams' crossed vowel conditions, this would indicate that focused attention to the pitch attributes of vowels instead of vowel color has little or no effect on categorization under these conditions. But, if the categorization shift in the SP-crossed condition is similar to the shift obtained in Williams' crossed nonspeech pitch condition, this result would indicate that it is attention to the perceptual dimension of pitch that influences how the stimuli are categorized regardless of whether the stimuli are speech or nonspeech.

The last possible result to be considered is that a categorization shift opposite in direction to the shifts obtained in Experiment 1 is observed in all four conditions. This result would be consistent with the hypothesis that auditory-based processes account for the perceptual effects of context.

Method

Subjects. Sixty-seven undergraduate students from introductory psychology courses at Indiana University volunteered to be subjects. Each subject took part in one 1-hour session and received partial course credit for participating in the experiment. All subjects were native speakers of English who reported no history of a speech or hearing disorder at the time of testing.

Stimulus Materials. The stimuli consisted of the same speech and sinewave tone-analog stimuli used in Experiment 1.

Procedure. Four experimental conditions were formed: Speech vowel crossed (SV-crossed), speech pitch crossed (SP-crossed), sinewave vowel crossed (SWV-crossed), and sinewave pitch crossed (SWP-crossed). As in Experiment 1, the experimental procedure used an AXB categorization task. However, in the present experiment, the "A" and "B" stimuli used in testing each continuum consisted of the endpoint stimuli from the opposite test series. For example, in the SV-crossed and SP-crossed conditions, the "A" and "B" stimuli for testing the steady-state [U]-[I] continuum consisted of the [wUw] and [wIw] endpoint stimuli from the time-varying continuum. Likewise, the "A" and "B" stimuli for testing the time-varying [wUw]-[wIw] continuum consisted of the steady-state [U] and [I] endpoints. This cross-series arrangement was carried out in a similar fashion for the steady-state and time-varying tone-analog stimuli in the SWV-crossed and SWP-crossed conditions.

All other aspects of the experiment were identical to Experiment 1. Stimulus presentation and data collection were controlled online by a PDP-11/34a computer.

Results

At the end of each experimental session, subjects completed the same post-test questionaire used in Experiment 1. Based on the same screening criteria as before, the data from 10 out of 10 subjects were analyzed for the SV-crossed condition, 12 out of 20 subjects were analyzed for the SWV-crossed condition, 12 out of 17 subjects were analyzed for the SP-crossed condition, and 12 out of 18 subjects were analyzed for the SWP-crossed condition.

The data were tabulated in terms of the percentage of times each stimulus in the continuum was judged to be like the "stimulus 1" endpoint. As in Experiment 1, the best-fitting normal ogives through these points was determined and the 50% points of the ogives used as the dependent measures entered into the statistical analyses.

Insert Figure 2 about here

Figure 2 displays the data collapsed over subjects for the steady-state and time-varying continua for the SV-crossed, SWV-crossed, SP-crossed, and SWP-crossed conditions. As shown in Figure 2, the categorization function for the time-varying stimuli appears to be shifted to the right of the categorization function for the steady-state stimuli in all four conditions. analyses.

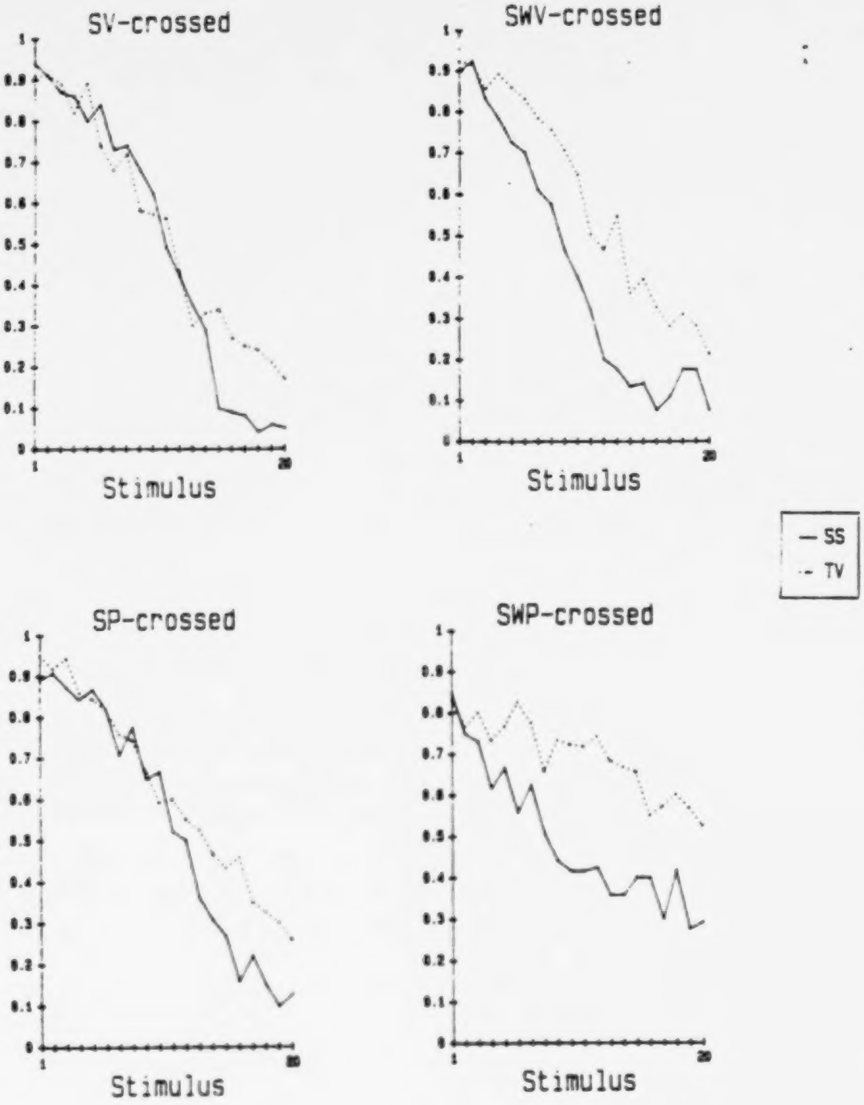


Figure 2. AXB categorization functions averaged over subjects for the steady-state [U]-[I] and context-dependent [wUw]-[wIw] continua for the SV-crossed, SWV-crossed, SP-crossed, and SWP-crossed conditions in Experiment 2.

Insert Table 2 about here

Table 2 shows the individual data in terms of mean crossover point values for all conditions. A separate one-way ANOVA on the data was conducted for each condition for the factor of continuum type. For the SV-crossed condition, a significant main effect of stimulus continuum was not obtained $F(1,11) = .5$, $p > .5$, although the mean crossover point value was greater for the time-varying stimuli compared to the steady-state stimuli (11.0 and 10.2, respectively). For the SWV-crossed condition, a significant main effect of continuum was obtained $F(1,11) = 9.8$, $p < .01$, with the mean crossover point value again greater for the time-varying stimuli compared to the steady-state stimuli (14.2 and 8.4, respectively). For the SP-crossed condition, a significant effect of continuum type was observed $F(1,11) = 4.9$, $p < .05$, with the mean crossover point value greater for the time-varying stimuli compared to the steady-state stimuli (14.6 and 11.3, respectively). Finally, for the SWP-crossed group, a significant main effect of continuum was not observed $F(1,11) = 2.0$, $p > .1$, with the mean crossover point value greater for the time-varying stimuli compared to the steady-state stimuli (76.0 and 7.8, respectively).

Thus, significant categorization shifts opposite in direction to the shifts we obtained in Experiment 1 were obtained for the SP-crossed and SWV-crossed conditions but not for the SV-crossed and SWP-crossed conditions. However, in both the SV-crossed and SWP-crossed conditions, the categorization shift was in the same direction as the other conditions. Upon examining the categorization function for the SWP-crossed condition in Figure 2 and the corresponding individual subject data in Table 2, it appears that the continuum endpoints for the SWP-crossed condition were not identified very accurately compared to the other conditions. Given that the data analysis for the majority of subjects in this condition yielded crossover point values that were beyond the range of the 20-item stimulus continuum, we decided to replicate this condition using a training procedure in order to ensure that the endpoint stimuli for the continuum were identified properly. In this manner, we could ensure that crossover point values within the range of the continuum could be obtained.

The training procedure that was instituted consisted of randomly presenting continuum stimuli with subjects receiving feedback about the correct AXB choice. First, 100 repetitions of the "stimulus 1" and "stimulus 20" endpoints were presented, for a total of 200 trials. Then, 50 repetitions of these two endpoints plus 50 repetitions of "stimulus 6" and "stimulus 15" were presented, for a total of 200 trials. Finally, after this preliminary training, all of the continuum stimuli were then presented for test without feedback.

Insert Figure 3 about here

Table 2

Individual subjects' category crossover point values for the steady-state (SS) and time-varying (TV) continua for all conditions in Experiment 2. The overall mean crossover values and standard deviations (slope values) for each condition are also displayed.

	SV-crossed		SWV-crossed		SP-crossed		SWP-crossed	
Ss	SS	TV	SS	TV	SS	TV	SS	TV
1	6.2	13.3	9.2	13.0	13.1	13.5	-36.0	29.3
2	7.5	4.0	9.4	14.1	8.7	9.4	11.4	14.3
3	13.2	15.2	7.7	12.3	9.9	14.1	-25.3	571.2
4	10.4	11.3	8.0	18.8	9.9	13.7	12.3	23.2
5	10.7	13.8	14.5	12.4	11.4	13.2	62.8	26.0
6	10.5	8.4	8.8	13.6	13.8	21.1	5.7	19.9
7	10.5	8.8	6.1	10.7	13.3	12.8	9.5	23.3
8	12.1	9.4	8.6	32.0	11.7	16.0	9.0	20.0
9	12.2	10.8	10.3	9.8	10.6	8.5	10.7	51.7
10	8.4	15.3	6.5	11.6	13.4	15.8	10.5	69.4
11	****	****	2.5	11.3	9.9	9.5	13.0	44.5
12	****	****	8.7	11.3	9.9	27.0	10.0	19.4
Mean	10.2	11.0	8.4	14.2	11.3	14.6	7.8	76.0
S.D.	-4.7	-10.4	-9.3	-15.2	-6.3	-12.4	-49.6	5.5

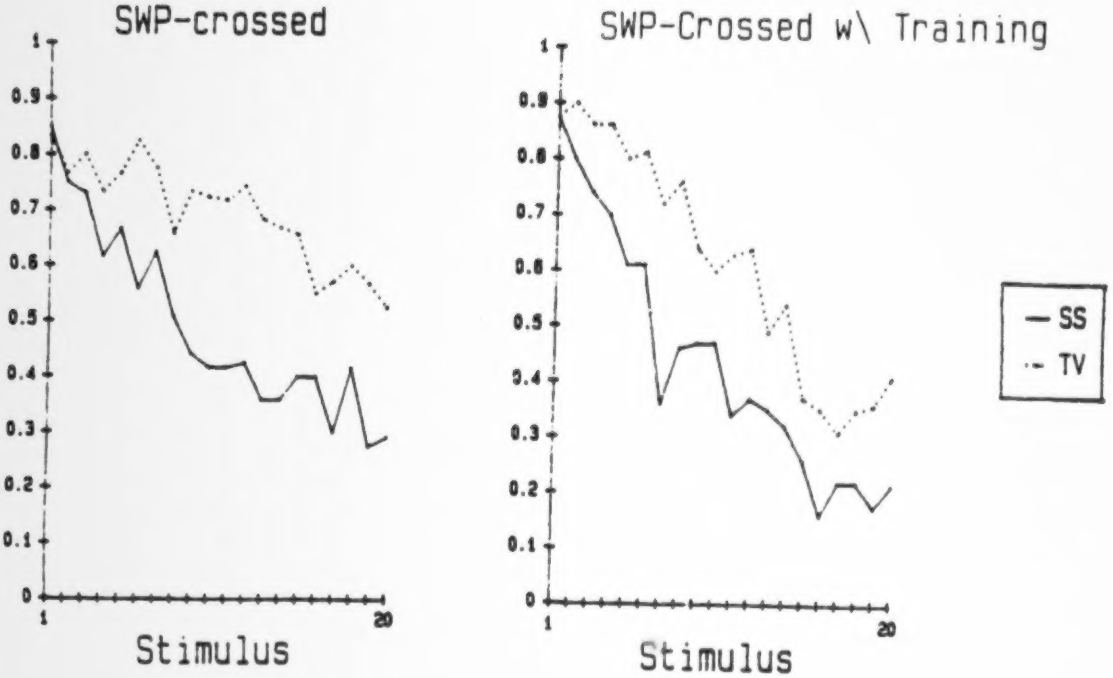


Figure 3. AXB categorization functions averaged over subjects for the steady-state [U]-[I] and context-dependent [wUw]-[wIw] continua for the SWP-crossed condition and the SWP-crossed replication with training in Experiment 2.

Figure 3 displays the categorization functions for the SWP-crossed condition and the replication of the SWP-crossed condition with training. The ANOVA showed that a significant effect of continuum was obtained for the SWP-crossed replication $\underline{F}(1,7) = 14.6$, $\underline{p} < .01$, with the mean crossover point value greater for the time-varying continuum compared to the steady-state continuum (15.8 and 8.7, repectively). Thus, a significant rightward categorization shift was observed as in the SP-crossed and SWV-crossed conditions.

Discussion

The results obtained in Experiment 2 differ markedly from the results we obtained in Experiment 1. First of all, in the SWP-crossed condition, we observed a significant shift in categorization responses in the opposite direction from that obtained in Experiment 1. This result replicates Williams' finding using similar nonspeech stimuli. This finding demonstrates that the perception of pitch in time-varying nonspeech stimuli is dependent on an integration of frequency information over the entire time-varying stimulus.

We also obtained a significant categorization shift in the SWV-crossed condition in the same direction as the shift in the SWP-crossed condition. This result demonstrates that vowel categorization and pitch categorization were affected in a similar manner by the cross-series manipulation. Thus, the perception of both pitch and vowel quality was significantly affected by the context in which the judgement is made. The similarity of the effects observed in the SWV-crossed condition and the effects observed in the SWP-crossed condition suggests that the processes or mechanisms operating on time-varying speech and nonspeech signals are very similar.

For the SP-crossed condition, a significant categorization shift was also obtained. The shift was in the same direction as the shifts obtained in the SWP-crossed and SWV-crossed conditions. This result indicates that, under crossed-endpoint conditions, categorization of vowels is similar regardless of whether vowel quality or pitch quality is used as the basis for the categorization judgement. Furthermore, this result demonstrates that the pattern of categorization is similar for both speech and nonspeech stimuli when pitch is used as the dimension of categorization. Thus, the categorization shift obtained in the SP-crossed condition provides evidence contrary to the conclusions of Williams regarding speech/nonspeech processing differences.

Finally, we failed to find a significant categorization shift in the SV-crossed condition. At first glance, this result seems inconsistent with both Williams' findings and with the results we obtained in the other three conditions. However, the direction of the shift in the SV-crossed condition was the same as in the other conditions, including the SP-crossed condition in which the vowel pitch of the speech stimuli was judged instead of vowel color. We have no apparent explanation for the discrepency between speech and nonspeech vowel conditions, other than to suggest that the context effects in the vowel conditions we have

73

observed in the present study may be somewhat less salient under certain conditions.

The results we obtained in Experiment 2 using the cross-series manipulation differ from those obtained by Williams (1986). The categorization shift in the SWP-crossed condition was similar in direction to the shift reported by Williams for his nonspeech pitch crossed condition. However, the categorization shift we observed in our SWV-crossed condition was in the opposite direction from the shift obtained in Williams' nonspeech vowel condition. Given that we used identical stimuli and identical AXB testing procedures, the reason for the difference in results between studies is not readily apparent. The only major difference between studies were that we used an explicit post-test screening procedure to ensure that subjects were processing the stimuli as speech or nonspeech. But, even if some of the subjects in Williams' nonspeech pitch crossed and nonspeech vowel crossed conditions heard the stimuli as speech or nonspeech, respectively, it is still difficult to reconcile the differences between studies. Thus, we can offer no reasonable explanation for the discrepencies observed between our study and Williams' study at this present time.

General Discussion

The results of the present investigation call into question the hypothesis that the perceptual processes involved in extracting vowel information from a coarticulated context are based on articulatory representations of speech. Given that perceptual compensation for time-varying context was observed for both speech and nonspeech stimuli across vowel and pitch categorization tasks, it seems reasonable to suppose that these results are due to general processes and/or mechanisms related to the early encoding of auditory signals. Given that the categorization of time-varying tone-analogs and coarticulated vowels was affected in approximately the same manner by manipulations to the AXB task, it appears that the perception of dynamic time-varying information contained in formant transitions leading into and out of the steady-state regions of vowel and vowel-like sounds is performed by processes that are integrative in nature. By integrative, we mean that computational-like processes integrate over the frequency or spectral information contained in the time-varying stimulus in order to produce pitch or vowel percepts, as the case may be. Given the pattern of results we obtained, we see no need for invoking processes related to articulatory knowledge or processes involved in a specialized "speech mode" of processing to explain these findings (Liberman et al., 1967; Liberman & Mattingly, 1985).

As noted earlier, Williams interpreted the absence of context effects in his sinewave pitch condition as evidence against an auditory-based account of vowel perception. Williams also interpreted the differences in categorization he observed between sinewave pitch and speech conditions in the cross-series experiment as support for two distinct perceptual mechanisms, one related to the auditory processing of nonspeech signals and one related to the processing of human speech. The results we obtained in the present study failed to replicate both findings reported by Williams. We believe that the present results raise questions about the major conclusions of his investigation. Our finding of context effects in Experiment 1 suggests that the processes involved in categorizing complex time-varying nonspeech signals and those involved in perceiving speech are probably very similar and engage similar types of perceptual mechanisms. In addition, our finding in Experiment 2 that the categorization of sinewave pitch and vowels is affected in the same manner by crossing the AXB endpoints provides further evidence for this conclusion.

An articulatory-based theory of speech perception would encounter severe problems in attempting to explain why perceptual compensation was observed with sinewave stimuli emulating speech, especially when post-test questionaires established that the listeners heard or processed the sounds as nonspeech tones. An articulatory-based theory would also have difficulty explaining why the perceptual effects produced by the cross-series manipulation were similar for both speech and nonspeech. Taken together, we believe that our results are most compatible with an auditory-based account of perceptual compensation. Although our results are consistent with an auditory-based explanation, an extension of these results to characterize all the perceptual processes utilized in the perception of fluent speech may be premature. Recently, a number of studies using nonspeech analogs of speech have shown differences in performance when the stimuli are heard as speech instead of nonspeech (Bailey et al., 1977; Best et al., 1981; Grunke & Pisoni, 1982; Ralston, 1986; Remez et al., 1981; Schwab, 1982; Tomiak et al., 1987). Results of this kind appear to present problems for purely auditory-based accounts of speech perception. However, it is entirely possible that general auditory processes perform some functions relevant to speech perception while other functions require the involvement of a cognitive component. By a cognitive component, we do not necessarily mean that certain functions in speech perception require the involvement of tacit knowledge of articulatory constraints or "intended gestures." For instance, it is possible that processing in a "speech mode" is related to cognitive and/or attentional factors that direct the perceptual system to process sequences of sounds as integral or separable perceptual units (see Tomiak et al., 1987). Such processes may be completely unrelated to articulatory knowledge or constraints. Alternatively, the processing of auditory signals in a "speech mode" may be related to attentional factors or strategies that weight the acoustic cues present in the speech signal in different ways (see Massaro, 1987; Oden & Massaro, 1978). This "weighting" may also have nothing to do with articulatory-related knowledge. The results obtained in the present study are fully compatible with postulating the presence of cognitive factors performing other functions in speech perception.

In summary, the present study provides new evidence suggesting that the processes employed in the perception of coarticulated vowels and dynamically time-varying sinewave patterns appear to be quite similar, despite a recent report by Williams. The contextual effects of consonantal formant transitions on vowel perception obtained by Lindblom and Studdert-Kennedy (1967) were replicated. The fact that we obtained effects of time-varying context in both speech and nonspeech categorization suggests that the conclusions drawn by Williams are not substantiated. Not only do our results argue that the perception of coarticulated vowels is not accomplished by an articulatory-based perceptual mechanism, they also argue that an articulatory-based explanation of speech perception leaves much to be desired. It would seem that a coherent articulatory-based theory accounting for speech perception in general should *always* require differences to be exhibited when performance between speech and nonspeech analogs of speech is examined. When the present findings are taken together with other research demonstrating that various perceptual phenomena found for speech are also found for nonspeech analogs of speech (Miller et al., 1976; Pisoni, 1977; Pisoni et al., 1983), one is compelled to seriously entertain the proposal that while there may be important differences in perception and sound generation between speech and nonspeech sounds, there also appear to be many similarities as well. Our results provide further support for the proposal that some aspects of speech perception are accomplished by general-purpose processes related to the encoding of auditory signals and the perception of complex auditory patterns.

References

Bailey, P.J., Summerfield, Q., & Dorman, M. (1977). On the identification of sine-wave analogues of certain speech sounds. *Haskins laboratories status report on speech research SR-51/52*, (pp. 1-25). New Haven: Haskins Laboratories.

Best, C.T., Morrongiello, B., & Robson, R. (1981). Perceptual equivalence of acoustic cues in speech and nonspeech perception. *Perception and Psychophysics*, **29**, 191-211.

Cole, R.A., & Scott, B. (1974). Toward a theory of speech perception. *Psychological Review*, **81**, 348-374.

Cutting, J.E. (1974). Two left-hemisphere mechanisms in speech perception. *Perception and Psychophysics*, **16**, 601-612.

Cutting, J.E. (1978). There may be nothing peculiar to perceiving in a speech mode. In J. Requin (Ed.), *Attention and performance VII*. Hillsdale, N.J.: Erlbaum.

Elman, J.L., & McClelland, J.L. (1986). Exploiting lawful variability in the speech wave. In J.S. Perkell and D.H. Klatt (Eds.), *Invariance and variability in speech processes*, (pp. 360-380). Hillsdale, N.J.: Erlbaum.

Fowler, C.A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics*, **14**, 3-28.

Grunke, M.E., & Pisoni, D.B. (1982). Some experiments on perceptual learning of mirror-image acoustic patterns. *Perception and Psychophysics*, **31**, 210-218.

Hillenbrand, J. (1984). Perception of sine-wave analogs of voice onset time stimuli. *Journal of the Acoustical Society of America*, **75**, 231-240.

Jenkins, J.J., Strange, W., & Edman, T.R. (1983). Identification of vowels in "vowelless" syllables. *Perception and Psychophysics*, **34**, 441-450.

Kewley-Port, D. (1976). A complex-tone generating program. *Research on speech perception progress report no. 3*. Bloomington, IN: Indiana University.

Klatt, D.H. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971-995.

Kuhl, P.K. (1986a). Theoretical contributions of tests on animals to the special-mechanisms debate in speech. *Experimental Biology*, **45**, 233-265.

Kuhl, P.K. (1986b). Reflections on infants' perception and representation in speech. In J.S. Perkell and D.H. Klatt (Eds.), *Invariance and variability in speech processes*, (pp. 19-30). Hillsdale, N.J.: Erlbaum.

Kuhl, P.K. (in press). On babies, birds, modules, and mechanisms: A comparative approach to the acquisition of vocal communication. In R.J. Dooling and S.H. Hulse (Eds.), *The comparative psychology of complex acoustic perception*. Hillsdale, N.J.: Erlbaum.

Liberman, A.M. (1970a). Some characteristics of perception in the speech mode. In D.A. Hamburg (Ed.), *Perception and its disorders, proceedings of A.R.N.M.D.* Baltimore, MD: Williams and Wilkins.

Liberman, A.M. (1970b). The grammars of speech and language. *Cognitive Psychology*, **1**, 301-323.

Liberman, A.M. (1982). On finding that speech is special. *American Psychologist*, **37**, 148-167.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological Review*, **74**, 431-461.

Liberman, A.M., Delattre, P.C., & Cooper, F.S. (1952). The role of selected stimulus variables in the perception of the unvoiced stop consonants. *American Journal of Psychology*, **52**, 127-137.

Liberman, A.M., Delattre, P.C., Cooper, F.S., & Gerstman, L.H. (1954). The role of consonant-vowel transitions in the perception of the stop and nasal consonants. *Psychological Monographs*, **68**, 1-13.

Liberman, A.M., Harris, J.S., Hoffman, H.A., & Griffith, B.C. (1957). The discrimination of relative-onset time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology*, **61**, 379-388.

Liberman, A.M., Isenberg, D., & Rakerd, B. (1981). Duplex perception of cues for stop consonants: Evidence for a phonetic mode. *Perception and Psychophysics*, **30**, 133-143.

Liberman, A.M., & Mattingly, I.G. (1985). The motor theory of speech perception revised. *Cognition*, **21**, 1-36.

Liberman, A.M., & Mattingly, I.G. (1989). A specialization for speech perception. *Science*, **243**, 489-494.

Lindblom, B.E.F. (1963). Spectrographic study of vowel reduction. *Journal of the Acoustical Society of America*, **35**, 1773-1781.

Lindblom, B.E.F., & Studdert-Kennedy, M. (1967). On the role of formant transitions in vowel recognition. *Journal of the Acoustical Society of America*, **42**, 830-843.

MacDonald, J., & McGurk, H. (1978). Visual influences on speech perception processes. *Perception and Psychophysics*, **24**, 253-257.

Massaro, D.W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry.* Hillsdale, N.J.: Erlbaum.

Mattingly, I.G., Liberman, A.M., Syrdal, A.K., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, **2**, 131-157.

McClelland, J.L., & Elman, J.L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, **18**, 1-86.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Miller, J.D., Wier, L., Pastore, R.E., Kelly, W., & Dooling, R. (1976). Discrimination and labeling of noise-buzz sequences with varying noise-lead times: An example of categorical perception. *Journal of the Acoustical Society of America*, **60**, 410-417.

Oden, G.C., & Massaro, D.W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.

Parker, E.M. (1988). Auditory constraints on the perception of voice-onset time: The influence of lower tone frequency on judgments of tone-onset simultaneity. *Journal of the Acoustical Society of America*, **83**, 1597-1607.

Parker, E.M., Kleunder, K.R., & Diehl, R.L. (1986). Trading relations in speech and nonspeech. *Perception and Psychophysics*, **39**, 129-142.

Pastore, R.E. (1981). Possible psychoacoustic factors in speech perception. In P.D. Eimas and J.L. Miller (Eds.), *Perspectives on the study of speech.* (pp. 165-206). Hillsdale, N.J.: Erlbaum.

Pastore, R.E., Harris, L.B., & Kaplan, J.K. (1982). Temporal order identification: Some parameter dependencies. *Journal of the Acoustical Society of America*, **71**, 430-436.

Pisoni, D.B. (1977). Identification and discrimination of the relative onset time of two component tones: Implications for voicing perception in stops. *Journal of the Acoustical Society of America*, **61**, 1352-1361.

Pisoni, D.B., Carrell, T.D., & Gans, S.J. (1983). Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, **34**, 314-322.

Pisoni, D.B., & Luce, P.A. (1987). Acoustic-phonetic representation in word recognition. In U.H. Frauenfelder and L.K. Tyler (Eds.), *Spoken word recognition*, (pp. 21-52). Cambridge, MA: MIT Press.

Ralston, J.V. (1986). *Auditory and phonetic perception of stop consonant place of articulation information.* Unpublished doctoral dissertation, State University of New York at Buffalo, Buffalo, N.Y.

Ralston, J.V., & Sawusch, J.R. (1984). *Perception of sinewave analogs of stop-consonant place information.* Paper presented at the 108th meeting of the Acoustical Society of America, Minneapolis, MN.

Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional cues. *Science*, **212**, 947-950.

Repp, B.H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, **92**, 81-110.

Repp, B.H. (1983). Trading relations among acoustic cues in speech perception: Speech-specific but not special. *Haskins laboratories status report on speech perception research SR-76*, (pp.129-132). New Haven, CT: Haskins Laboratories.

Repp, B.H. (1984). Against a role of "chirp" identification in duplex perception. *Perception and Psychophysics*, **35**, 89-93.

Schouten, M.E.H. (1980). The case against a speech mode of perception. *Acta Psychologica*, **44**, 71-98.

Schwab, E.C. (1982). Auditory and phonetic processing for tone analogs of speech. *Dissertation Abstracts International*, **42**, 3853B.

Stevens, K.N., & Halle, M. (1967). Remarks on analysis by synthesis and distinctive features. In W. Wathen-Dunn (Ed.), *Models for the perception of speech and visual form*, (pp. 88-102). Cambridge, MA: MIT Press.

Stevens, K.N., & House, A.S. (1963). Perturbation of vowel articulations by consonantal context: An acoustical study. *Journal of Speech and Hearing Research*, **6**, 111-128.

Strange, W. (1987). Information for vowels in formant transitions. *Journal of Memory and Language*, **26**, 550-557.

Strange, W., Jenkins, J.J., & Johnson, T.L. (1983). Dynamic specification of coarticulated vowels. *Journal of the Acoustical Society of America*, **74**, 695-705.

Studdert-Kennedy, M., & Shankweiler, D.P. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America*, **48**, 570-594.

Studdert-Kennedy, M., Liberman, A.M., Harris, K.S., & Cooper, F.S. (1970). Motor theory of preparation: A reply to Lane's critical review. *Psychological Review*, **77**, 234-249.

Summerfield, Q. (1979). Use of visual information for phonetic perception. *Phonetica*, **36**, 314-331.

Tomiak, G.R., Mullennix, J.W., & Sawusch, J.R. (1987). Integral processing of phonemes: Evidence for a phonetic mode of perception. *The Journal of the Acoustical Society of America*, **81**, 755-764.

Whalen, D.H., & Liberman, A.M. (1987). Speech perception takes precedence over nonspeech perception. *Science*, **237**, 169-171.

Williams, D.R. (1986). *Role of dynamic information in the perception of coarticulated vowels.* Unpublished doctoral dissertation, University of Connecticut, CT.

Woodworth, R.S. (1938). *Experimental psychology.* New York: Holt.

81

Appendix A

Post-test Information Questionaire

We are interested in your reactions to the stimuli that you listened to during the experiment today. Please answer the following questions as thoughtfully as possible. Your responses to these questions will be very helpful to us in our research program on auditory perception here in the Psychology Department. Please circle the appropriate yes or no answer for each question. Thank you.

1. The stimuli sounded like sounds yelping dogs might make. Yes No
2. The stimuli sounded like they could be used as voices in a computer video game. Yes No
3. The stimuli sounded like vowels that a talking robot might say. Yes No
4. The stimuli sounded like tones from a musical synthesizer. Yes No
5. I have never heard sounds like these before this experiment. Yes No
6. I have heard these sounds before in a previous experiment in this or another lab. Yes No
7. The stimuli sounded like good approximations to human vowel sounds. Yes No
8. The stimuli sounded like vowels, but of poor quality compared to human vowels. Yes No
9. The stimui sounded like they came from a wind-up doll. Yes No
10. I made my responses to the stimuli by matching the pitch of the second sound to the first and third sounds. Yes No
11. I made my responses to the stimuli by matching the vowel in the second sound to the vowel in the first and third sounds. Yes No
12. The stimuli sounded like tones that you get in tests of your hearing. Yes No
13. When listening to the stimuli, I did not hear them as vowels at all. Yes No
14. The sounds I heard were very unnatural, but they did sound like vowels to me. Yes No
15. The sounds I heard sounded like they were vowels from a foreign language. Yes No

Intonational Context and F0 Normalization[1]

Keith Johnson

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] **I appreciate the comments that I have received from Mary Beckman, Rob Fox, Ilse Lehiste, Neal Johnson, John Mullennix, David Pisoni and Van Summers. The research reported here was supported by NIH Training Grant *NS 07134-10* to Indiana University in Bloomington, IN.**

Abstract

A two-stage model of vowel normalization is proposed and tested. In the first stage, the harmonics in the region of the first oral resonance (F1) and fundamental frequency (F0) are auditorily integrated. In the second stage, information concerning estimated formant ranges for the speaker is utilized in evaluating formant values. A process of range estimation is proposed in which formant ranges are estimated upon the basis of cues to speaker-identity both internal to the vowel (F0) and in the phonetic context of the utterance. In Experiment 1, subjects identified words from two "hood"-"hud" continua - one with low F0 (100 Hz) and one with high F0 (150Hz). The words were presented in two conditions - in isolation and embedded in one of two different carrier phrases. In a pretest it was found that the intonation pattern of the carrier phrases virtually eliminated the speaker differences for the two levels of F0 while retaining the fundamental frequency differences that the words had in isolation. There was a shift in identification as a result of the change in F0 for both the words presented in isolation and for the words presented in intonational contexts, however, the shift was not as great when items were presented in context. Experiment 2 was conducted as a test of the possibility that presentation in a carrier phrase influences vowel normalization, regardless of the degree of speaker difference from one context to the other. Tokens from the "hood"-"hud" continuum were again presented in isolation and in carrier phrases. Again, there were two levels of F0 (130 Hz and 150 Hz) and two types of intonational contour. In this experiment the high and low F0 items in carrier phrases and in isolation had the same degree of speaker-identity difference (as determined in the pretest). The effect of F0 upon vowel identification in this experiment was of the same magnitude in both the context condition and the isolation condition. This was taken to mean that the results of Experiment 1 were due to the speaker-identity differences of the tokens and not merely to an effect of being presented in a carrier phrase. Experiment 3 involved the presentation of vowels from the "hood"-"hud" continuum in two different intonational contexts which were judged to have been produced by different speakers even though the F0 of the test word was identical in the two contexts. There was a shift in identification as a result of the intonational context which was interpreted as evidence for the role of speaker identity in vowel normalization.

Intonational Context and F0 Normalization

Of all of the hearer's accomplishments in perceiving speech, one of the most important is the ability to adjust to different talkers. In this paper I report the results of three experiments which examine the role of F0 in vowel perception. Discovering how it is that hearers use F0 in the processes involved in vowel perception will provide important clues concerning the ways in which hearers adjust to different talkers. There are two kinds of computational models of vowel normalization: range normalization and F0 normalization. There is also perceptual evidence which suggests that hearers utilize perceptual processes analogous to both of these computational approaches in normalizing for speaker differences.

Adjusting to Talkers: Two Approaches

One computational technique for classifying vowels produced by different talkers is called range normalization (Gerstman, 1968; also see Nearey, 1978).[1] Gerstman's (1968) method for performing range normalization involves scaling vowel formant measurements for a particular speaker to some standard scale. In order to perform this scaling operation, it is necessary to know the *range* of values which are possible for the particular speaker. In the example used by Gerstman, the minimum and maximum values of each formant for each speaker in the Peterson and Barney (1952) database of vowel formant measurements served to define the formant ranges for each speaker. Once the ranges for the formants for a particular speaker have been determined, it is possible to express the formant measurements relatively rather than absolutely. So, for instance, if we find that the F1 measurements for a particular speaker range from 300 Hz to 600 Hz then the range normalized value for an F1 of 450 Hz will be 0.5 (with 300 Hz being 0.0 and 600 Hz being 1.0). By converting the formant measurements to range values, Gerstman found that the differences between individual speakers were greatly reduced while vowel distinctions were maintained.

As an hypothesis about speaker normalization in perception this approach suffers an obvious problem. Hearers do not have any exact knowledge of the formant ranges for each new person that they listen to, and more importantly, hearers seem to perform vowel identification perfectly well even when the identity of the speaker is unpredictable from trial to trial (Verbrugge et al., 1976). Gerstman (1968) suggested that in order for a range normalization algorithm to work, the speech to be identified should be preceeded by examples of the corner vowels [i], [u], and [a] as a means of establishing the formant ranges for a particular speaker. The fact that hearers do not need such range-setting tokens is an indication of the inadequacy of the approach as a model of human vowel normalization.

Having said this, let's look at some evidence which seems to indicate that range normalization *does* take place in perception. Ladefoged and Broadbent (1957) found that subjects' identifications of test vowels could be influenced by the vowel formant values of a preceeding context. This influence is just what one would expect if the hearer performs some sort of range normalization. The vowel formants in the preceding context provide for the hearer an indication of the range of the speaker's vowel formant values (which are determined to a large extent by the length of the speaker's vocal tract, the longer the vocal tract the lower the vowel resonances). Thus, the expected range of the speaker's formant values influences the hearer's identification of the test vowel. This has been called "vocal tract normalization" because it seems that the hearer is evaluating the vowel formants by reference to the perceived length of the speaker's vocal tract. Vocal tract normalization has been demonstrated in a number of other studies using natural speech (van Bergem, et al., 1988), isolated point vowels instead of a context phrase (Ainsworth, 1975, Nearey, 1978), and sinewave analogs of speech (Remez, et al., 1987).

---

[1] **Nearey's (1978) constant ratio hypothesis is essentially a range normalization procedure in which the range is determined by a single vowel (hence "point normalization").**

From these studies, it is clear that hearers can estimate a speaker's formant ranges by reference to the formant values of vowels in a preceding context. Formant range information need not be supplied by preceding context, however. It is possible that hearers use vowel internal information to estimate the ranges of a speaker's formants. Although it is not possible to determine a speaker's formant ranges exactly from a single syllable, there are clues to be found within the vowel that can give some indication of the speaker's characteristic formant ranges. By using F0 and perhaps glottal source characteristics as an indication of vocal tract size, it may be possible for the hearer to estimate the speaker's formant ranges from phonetic information within an isolated syllable. Thus, a type of range normalization may take place in the perceptual processing of isolated vowels.

The second basic approach to modelling vowel identification behavior has also made use of F0 information, but in a quite different way. The work of Syrdal and Gopal (1986) is representative (see also Miller, 1987; and Hillenbrand and Gayvert, 1987). In this approach, vowel classifications are not determined by the absolute values of the vowel formants, but rather by the relative values of the formants and F0. Thus, the traditional vowel feature "openness" is determined by the F1/F0 ratio, the feature "compactness" is indicated by the F2/F1 ratio, and the F3/F2 ratio is used to determine whether the vowel is "front" or "back". Clearly, in this strategy, each vowel token contains all of the information necessary to perform normalization. In this way, it is possible to perform speaker normalization for a given token without explicitly identifying vowel formant ranges as is necessary in a range normalization procedure. Since it has been suggested that the harmonics in the F1 region and F0 may be auditorily integrated (Traunmüller, 1981; Sussman, 1986). , I will call this approach "auditory F0 normalization".

As with the range normalization approach, there is perceptual evidence that seems to be consistent with the auditory F0 normalization model. Chiba and Kajiyama (1941) concluded from a study of the perception of vowels which were reproduced at varying playback speeds that the ratios of the formants are the crucial cues for vowel identity (as opposed to the absolute locations of spectral peaks). This formant ratio hypothesis was also suggested by Potter and Steinberg (1950) and more recently by Nearey (1978). Potter and Steinberg suggested that "within limits, a certain spatial pattern of stimulation along the basilar membrane may be identified as a given sound regardless of position along the membrane"(p. 812). Thus, speaker normalization might have an explanation in terms of the auditory coding of vowels independent from any higher-level "adjustment to talker" process.

Traunmüller (1981) found that, in the identification of single-formant vowels, the perception of openness seemed to depend upon the ratio of F1 to F0. Shifts in degree of openness occurred at about the same value of F1/F0 ratio for vowels with F0 from 150 to 350 Hz. In a second experiment he demonstrated that shifting F0 and F1 (in five-formant vowels) equally along a modified tonality scale did not greatly change perceived openness, while shifting F1 alone (in a third experiment) resulted in category shifts. Finally, he varied F0 alone and found a shift in perceived openness similar to findings reported by Miller (1953), Fujisaki and Kawashima (1968) and Slawson (1968). Unlike previous researchers, however, Traunmüller explains the F0 normalization process as auditory in nature rather than as a higher-level perceptual process.

Traunmüller's description of the auditory process of F0 normalization makes reference to the "centre of gravity" effect reported by Chistovich et al. (1979). They found that when the interval between two formants (in two-formant vowels) is less than 3 to 3.5 Bark, a single-formant vowel which is judged to match the quality of the two-formant standard has F^* (the single-formant) such that $F1<F^*<F2$. However, when the interval between F1 and F2 exceeds 3 Bark, F^* is no longer positioned between the two formants of the standard but rather is adjusted to match the F2 of the standard. Chistovich et al. hypothesized that this pattern of response is the result of an auditory process. They proposed that vowel perception involves a second frequency integration process, after the first integration over critical band intervals (see Zwicker, 1961 and Scharf, 1970). This second stage of spectral smoothing is hypothesized to extend over a range of 3 to 3.5 Bark. Traunmüller (1981) hypothesized that the role of F0 in the perception of vowels arises from the integration of the lowest harmonics and the harmonics around the first formant into one "centre of gravity". Thus, as F0 increases and the lowest harmonics move closer to the F1 region, they play a greater role in the estimation of the spectral peak for F1. As a result, the perceived F1 is drawn down toward F0 even when the actual resonance of the vocal tract remains fixed.

One crucial difference between the range normalization approach and the auditory F0 normalization approach is that the first makes reference to the identity of the speaker while the second does not. Thus, according to the range normalization approach, if a hearer identifies two utterances as having been produced by different talkers (who have different vocal tract sizes), then the perceived quality of vowels which have identical F0 and identical formant values will be different because of the different expected formant ranges. Alternatively, the auditory F0 normalization hypothesis predicts that if F0 and F1 are the same, the perceived vowel quality will not be affected by speaker differences. Of course, it is also possible that both auditory F0 normalization and range normalization take place (Slawson, 1968). The model that seems to be indicated by the perceptual literature reviewed above is one that involves two stages of normalization: first, an auditory F0 normalization process which results from the auditory integration of F0 and F1, and second, a range normalization process in which vowel formants are interpreted by reference to perceived speaker identity (perceived vocal tract length). It is also possible that the estimation of a speaker's formant ranges may be accomplished both by reference to preceding vocalic context and vowel internal speaker information (particularly F0).

This dual-process view of normalization fits into a more general view of speech perception that includes an auditory stage of **information extraction** and a phonetic stage of **information integration**. At the auditory stage, basic features of the signal such as F0, formant peaks[2] and vowel duration are extracted. It is hypothesized that the process of formant peak extraction involves a smoothing of the frequency spectrum along the lines suggested by Chistovich et al. (1979). F0 may influence perceived F1 at this stage of processing by playing a role in the "centre of gravity" for F1 as proposed by Traunmüller (1981).

The phonetic stage of the model consists of processes of information integration. At this stage, various basic features are interpreted by reference to other basic signal dimensions (e.g. vowel quality by reference to vowel duration) or by reference to situational dimensions (e.g. vowel duration by reference to speaking rate). In the perception of isolated syllables an estimate of the situational dimension "range of formant values" (or "length of speaker's vocal tract") may be derived from the fundamental frequency of voice.[3]

Models that take into account both auditory and phonetic levels of processing, such as the one proposed here for vowel normalization, have proven useful in the description of consonant perception (Sawusch, 1977; Sawusch and Nusbaum, 1983) and vowel perception (Sawusch, Nusbaum and Schwab, 1983; Fox, 1985). There is no reason to believe that vowel normalization does not also involve both auditory and phonetic levels of processing.

The fact that there is perceptual evidence for both auditory F0 normalization and range normalization seems to require a two-stage model. However, the evidence which has been taken to support the auditory F0 normalization approach confounds F0 information with speaker information. In all of the studies that have demonstrated an F0 normalization effect (Miller, 1953; Fujisaki and Kawashima, 1968; Slawson, 1968 and Traunmüller, 1981), the vowel sounds to be identified have been presented in isolated syllables with different levels of F0. Since the only stimulus variable which changes from one condition to the next is F0, it is natural to talk about changes in perceived vowel quality as a function of F0. It seems likely, however, that confounded with the differences in F0 were also differences in perceived speaker quality, because F0 is a strong cue for speaker identity (Carrell, 1984). Thus, it would be possible, from a range normalization point of view, to consider the changes in perceived vowel quality to be a function of speaker identity.

---

[2] Whether or not formant peaks are important in vowel perception is a controversial issue. Those in favor of a whole spectrum approach include Papçun (1980), Bladon (1982), Plomp (1975). Evidence for the importance of formant peaks as opposed to whole spectrum representation is provided by Chistovich (1985), and Assman and Summerfield (in press).

[3] Clearly, speaker identity depends upon more than just F0, however, when formant values are ambiguous, F0 may be the primary cue (see Carrell, 1984).

In the experiments reported here, I attempt to separate (1) changes in vowel quality due to changes in F0 and (2) changes in vowel quality are due to changes in speaker characteristics. To do this, I embedded test words in carrier phrases. Manipulation of the intonational contours of the phrases allowes for a certain degree of control over perceived speaker identity. This manipulation makes use of the fact that, in natural settings, hearers are confronted (1) with speech from the same talker in which F0 is variable and (2) with speech from different talkers in which F0 ranges overlap.

Experiment 1 involved the presentation of tokens from a "hood"-"hud" continuum in carrier phrases with intonational contours that sounded like they had been produced by the same talker, even though the F0 of the test words was, in one contour, 150 Hz and, in the other, 100 Hz. The experiment thus avoided confounding speaker identity with F0. Experiment 2 addresses the question of whether simply embedding test words in an intonational context affects vowel identification behavior indepenently from the F0 or speaker-identity of the tokens. Identification functions for vowels in a carrier phrase and in isolation were compared when both the context and the isolated words were equated for speaker-identity and F0. Experiment 3 involved the presentation of the "hood"-"hud" continuum in carrier phrases which were judged to have been produced by different talkers while the F0 of the test word was identical in the two phrases.

Experiment 1

In Experiment 1, the F0 of synthesized test tokens along a continuum from [hʊd] to [hʌd] was set at 100 Hz for one version of the continuum and 150 Hz for another. The test tokens were presented in a forced-choice identification task either in isolation or as the last word in one of two carrier phrases. The phrase "this is hVd" was synthesized with a rising intonational contour (question intonation) or a falling intonational contour (statement intonation). The 150 Hz tokens were embedded in the rising intonational context and the 100 Hz tokens were embedded in the falling intonational context. The two carrier phrases were judged in a pretest (see Appendix) to have been produced by the same speaker. Isolated test words with F0 of 100 Hz and 150 Hz were classified in the pretest as having been produced by different speakers. Use of intonational contexts made it possible to separate speaker-identity from F0.[4] The items in isolation differed both in F0 and in speaker-identity and therefore any shift in vowel identification from high F0 to low F0 could be due to both auditory F0 normalization and range normalization. The items presented in an intonational context differed in F0 but not in speaker identity, therefore these tokens provide a test of the auditory explanation of F0 normalization in which F0 is not confounded with speaker-identity. This experiment is also a test for the existence of a range normalization process because the only difference between the items in isolation and the items in intonational context is one of speaker-identity. If there is a range normalization process that makes reference to speaker-identity, I predict that there will be a greater shift in vowel identification for items presented in isolation than for items presented in context.

---

[4] In using the term speaker-identity I do not mean to imply that the hearers attributed the same speaker

Method

Subjects. Twenty undergraduate psychology students at Indiana University (5 males, 15 females) served as subjects in the experiment. They received partial course credit for their participation. All subjects reported no history of speech or hearing disorder, were naive to the purpose of the experiment and participated in a single one hour session.

Materials. A seven step [hæd]-[hʌd] continuum was synthesized once with F0 of 100 Hz and again with F0 at 150 Hz. Formant values of the vowels in the continuum are given in Table 1. The endpoints of the vowel continuum had formant values between the average values reported for male [ʌ] and female [æ] as reported by Peterson and Barney (1952) and so were ambiguous with regard to speaker sex. The steps in the continuum were equally spaced in Bark.

Insert Table 1 about here

The [hVd] words were synthesized in isolation and in the intonational contexts indicated in Figure A1 in the Appendix. The 150 Hz tokens were presented in the rising intonational contour which was used as the standard in the pretest while the 100 Hz tokens were synthesized in the falling intonational context which in the pretest was judged to be most similar in speaker-identity to the standard (75% "same" judgements). The tokens were steady-state vowels (200 ms) in the consonantal environment [hVd]. Duration of the /h/ portion was 60 ms. The final transitions to /d/ had a duration of 30 ms. Overall duration of the test words was 290 ms. As with the vowel formants of the test tokens, the formant values for [ι] in the carrier phrase ("This is ____") were chosen so as to be ambiguous between values typical of males and females (F1 = 456, F2 = 1740, F3 = 2796). "This" was 415 ms long, with a 210

---

qualities to the voices they heard (such as "tall", "heavy", etc.) I merely mean that in the pretest subjects classified two synthesized phrases or words as having been produced by the same speaker, or by different speakers. This operational definition is adequate for the purpose at hand.

Table 1

Formant values of the test tokens used in Experiments 1-3.

Token #	1	2	3	4	5	6	7
F1	474	491	509	526	543	561	578
F2	1111	1124	1137	1150	1163	1176	1189
F3	2416	2424	2432	2440	2448	2456	2464

ms steady-state [ɩ]. "Is" was 195 ms long, 90 ms of which was the vowel. Overall duration of the phrase (including the test word) was 900 ms.

Procedure. The experiment was conducted on-line using a PDP 11/34 computer at the Speech Research Laboratory at Indiana University. Tokens were played at 10 kHz, low pass filtered at 4.8 kHz, and presented binaurally over matched and calibrated TDH-39 headphones at a peak SPL level of 85 dB. Subjects were run in groups of up to six at a time, and reaction time as well as response data were collected. Labels for the response buttons used by subjects were displayed on a CRT at eye level and the association between response categories and buttons was switched on successive blocks.

Half of the subjects heard the isolated tokens and half responded to the tokens in intonational context. Each token was presented 10 times and the order of presentation was randomized.

A between subjects design was used with presentation type as a grouping variable. Each subject responded to 140 presentations (7 tokens X 2 F0 levels X 10 repetitions).

Results and Discussion[5]

The vowel identification data (Figure 1) were analysed in a repeated-measures analysis of variance. The between subjects factor was PRESENTATION TYPE (items in isolation vs. items in intonational context), and the within subjects factors were: TEST WORD F0 (high = 150 Hz vs. low = 100 Hz) and TOKEN NUMBER (see Table 1). The TOKEN NUMBER and F0 main effects were significant [$F(6, 108) = 68.53$ and $F(1, 18) = 39.06$, respectively, both $p < 0.001$]. The TOKEN X PRESENTATION TYPE and TOKEN X F0 interactions were significant [$F(6, 108) = 5.15$ and $F(6, 108) = 4.87$, respectively, both $p < 0.001$]. As can be seen in Figure 1, PRESENTATION TYPE and F0 level affected the ambiguous middle tokens in the continuum more than they did the continuum endpoints. It is also clear from Figure 1 that the F0 effect was greater when items were presented in isolation than when the test words were presented in context. This interaction (PRESENTATION TYPE X F0) was significant [$F(1, 18) = 10.04, p < 0.01$]. The three way interaction

---

[5] Reaction time data were collected, but they do not particularly bear on the hypothesis under consideration. A brief discussion of the general pattern of this data will be given here. This pattern was observed in all three experiments in this study. The reaction time data for the isolated tokens conformed with the pattern found by Pisoni and Tash (1974) (ambigous items were identified more slowly than unambigous items). The reaction times to the items in context were on average faster than to the items in isolation and were undifferentiated across the continuum (no difference between ambiguous and unambiguous items). This conforms to the prediction of Egan (1948) that context serves to allow subjects to prepare for a response. The lack of differentiation between items in the continuum can then be interpreted as a floor effect. Since these considerations do not help in separating the hypotheses under consideration no further analysis of reaction time data was performed.

(F0 X PRESENTATION TYPE X TOKEN) was not significant [$F = 1.37$].

Insert Figure 1 about here

Percent "hood" responses averaged across all tokens are presented in Table 2. The values in this table make it possible to estimate the relative magnitudes of the two perceptual processes (auditory F0 normalization and range normalization). An auditory F0 normalization effect can estimated from the degree of shift that remains when speaker differences have been eliminated. This is the value for Δ HL in the context condition (a difference of 17.86% "hood" responses). Note, however, that in the pretest these two contexts were not judged to be the "same" speaker 100% of the time, so part of this difference between the two levels of F0 may be some residue of speaker-identity difference. The range normalization effect can be estimated by the degree of difference between the shift for items in isolation versus the shift for items in context (ΔΔ HL) in Table 2. These data indicate that the shift in vowel identification when F0 is changed from 150 Hz to 100 Hz for tokens presented in isolation can be divided into auditory F0 normalization and range normalization at a ratio of about 1/2 (17.86/36.7). In other words, about 1/3 of the shift can be attributed to auditory F0 normalization and about 2/3 of the shift is due to range normalization.[6]

Insert Table 2 about here

In order to compare the auditory F0 normalization effect observed here with the F0 normalization effects reported elsewhere, the degree of boundary shift in the context condition was calculated in terms of %F1 shift per doubling of F0. Fifty percent crossover boundaries were calculated by linear interpolation from the identification functions in Figure 1. The boundaries corresponded to F1 values of 511 and 536 Hz for the low and high F0 continua, respectively. The %F1 shift per 100% shift in F0 (calculated by formula (1)) was 9.8%. This is on the low end of the 10% to 20% per doubling F0 reported by Nearey (1987). This is consistent with the hypothesis that the F0 normalization effects reported previously include both an auditory and a phonetic component. I would expect a lower estimate of the role of F0 in vowel identification when F0 is separated from speaker-identity as has been done here.

(1) $((\Delta F1/F1_{low})/(\Delta F0/F0_{low})) * 100$
$F1_{low}$ is the lower of two F1 boundaries and $F0_{low}$ is the lower of the two F0 values used as independent variables in the experiment.

---

[6] As will be demonstrated in Experiment 3, these estimates of the effects cannot be taken as absolutes. In particular, the magnitude of the F0 effect for the isolated condition is probably increased by a contrast effect (Johnson, 1988).

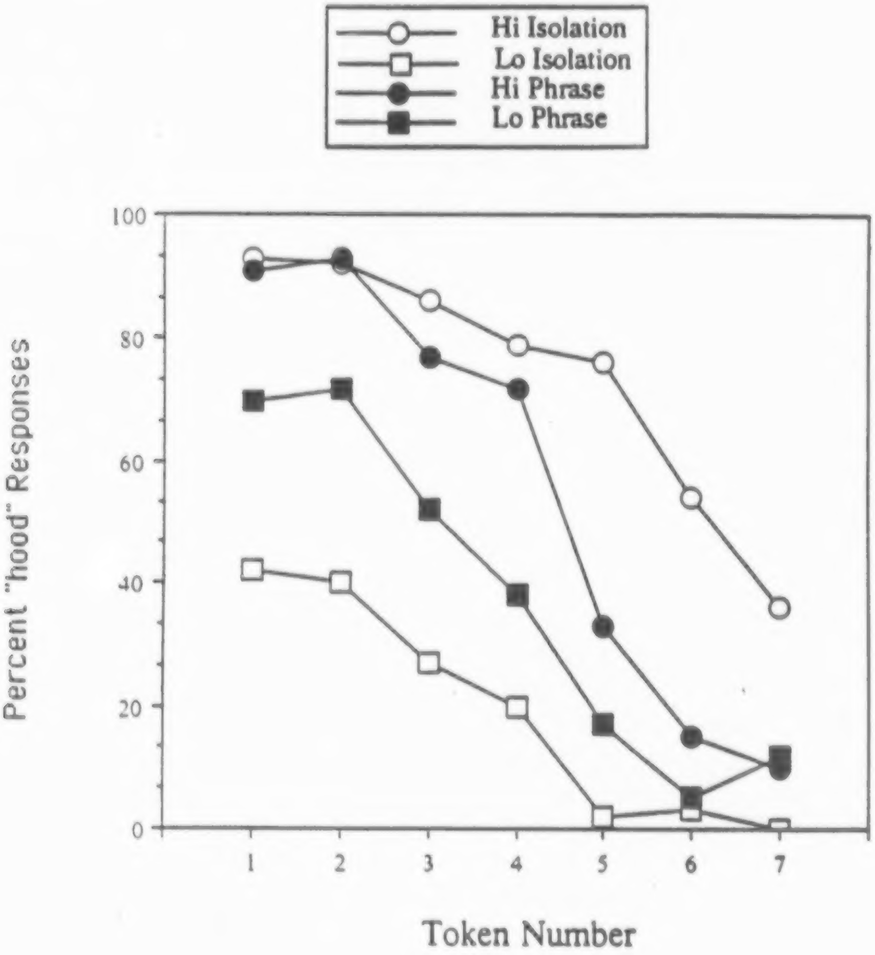


Figure 1. Results of Experiment 1. Identification functions for both levels of F0 when the tokens were presented in isolation versus when they were presented in an intonational context.

Table 2

Results of Experiment 1. Percent "hood" responses are shown for the high and low F0 tokens in isolation and in an intonational context. Δ HL is the difference between the %"hood" responses to the two F0 levels and ΔΔ HL is the difference of the differences.

F0 =	150 Hz	100 Hz	Δ HL
Isolation	73.70	19.14	54.56
Context	55.86	38.00	17.86
		ΔΔ HL	36.70

It is possible, though, that the effects for presentation type found here involve something other than speaker-identity. It may be that the perception of a vowel continuum will be constrained by the presence of a preceding context in ways other than the simple speaker-identity measure identified in the pretest. To investigate the possibility of the presence of some unforeseen confounding influence, an experiment was conducted in which the speaker-identity scores and the F0 values of isolated tokens and intonational phrases were matched. If there is some effect for presentation in a context in addition to these two factors (F0 and speaker-identity), I would expect this experiment to reveal a difference for these matched phrases and words.

Experiment 2

In Experiment 2, subjects identified vowels from "hood"-"hud" continua which were synthesized at steady-state F0 levels of 130 and 150 Hz. The tokens were also presented in intonational contexts which were classified in the pretest as having been produced by the same speaker 40% of the time. The tokens in isolation were classified as having been produced by the same speaker 51% of the time. Thus, in this experiment I expect the contribution of speaker-identity difference to be roughly equal for the two presentation conditions, and of course, I expect the contribution of F0 differences to be the same whether the tokens are presented in isolation or in a phrase.

Based on the results of Experiment 1 the following predictions are possible: (1) since speaker-identities are roughly equal in the two presentation conditions we predict that the degree of shift due to changing F0 will be roughly the same whether the items are presented in an intonational context or in isolation and (2) since both presentation types include speaker differences as well as F0 differences we predict that the degree of shift in F1 boundary, as a function of F0 doubling, will be greater for these data than was the shift in F1 boundary for the items in context in Experiment 1 (where speaker differences were reduced to a minimum).

Method

Materials and procedure were identical to those in Experiment 1, the only change being that instead of tokens synthesized at 100 and 150 Hz (with their respective intonational contours from the pretest) the tokens used in this experiment were synthesized with F0 of 130 and 150 Hz. The subjects (8 males and 12 females) were native American English speaking students at Indiana University. They reported no history of speech or hearing problems. Fifteen of the subjects received partial course credit for their participation and five were paid a small sum for participating.

Results and Discussion

Data were submitted to a repeated measures ANOVA with between subjects factor: PRESENTATION TYPE (isolated or in context) and within subjects factors: TEST WORD F0 (high or low) and TOKEN NUMBER. There were ten subjects in each group. There was a main effect for TOKEN NUMBER $[F(6,108) = 157.48, p < 0.001]$. The F0 main effect and the F0 X TOKEN interaction were also significant $[F(1,18) = 13.82, p < 0.01]$ and $[F(6,108) = 5.42, p < 0.001]$, respectively. As illustrated in Figure 2, the F0 X TOKEN interaction occurred because there was a boundary shift in response to the change in F0 level. Note also that, as predicted, presentation type did not play a role in subjects' identification responses.

Insert Figure 2 about here

The fifty percent crossover boundaries on these functions correspond to values of F1 of 535.8 and 517.9 Hz for the high and low F0 continua, respectively. Expressed relative to the degree of F0 shift in this experiment, this F1 shift corresponds to a 22.5% increase in F1 boundary per doubling of F0. This value is markedly greater than the degree of shift found for the items in carrier phrases in Experiment 1 and conforms to the predictions of a model of vowel normalization in which both F0 level and speaker identity play a role.

The fact that there was no context effect in this experiment indicates that the data in Experiment 1 for the items presented in carrier phrases reflect the operation of a process sensitive to speaker-identity, not simply an effect of presentation type (items in isolation versus items in a carrier phrase).

Experiment 3

In this final experiment, subjects identified vowels from the "hood"-"hud" continuum that were synthesized in carrier phrases which had intonational contours such that the items sounded like they had been produced by different speakers (a high-pitched voice making a statement and a low-pitched voice asking a question). The characteristic of the carrier phrases which is of most interest here is that they had identical test-word F0. The situation created by the use of these carrier phrases is one in which speaker-identity varies as a result of the F0 of the carrier phrase while test word F0 is held constant. If vowel normalization involves only an auditory F0 normalization process I would expect to find no difference in vowel identification as a result of placing the test words in these contexts. If, on the other hand, there is a process of range normalization (which involves the use of F0 to estimate the vowel formant range) I would predict that there will be a shift in vowel identification such

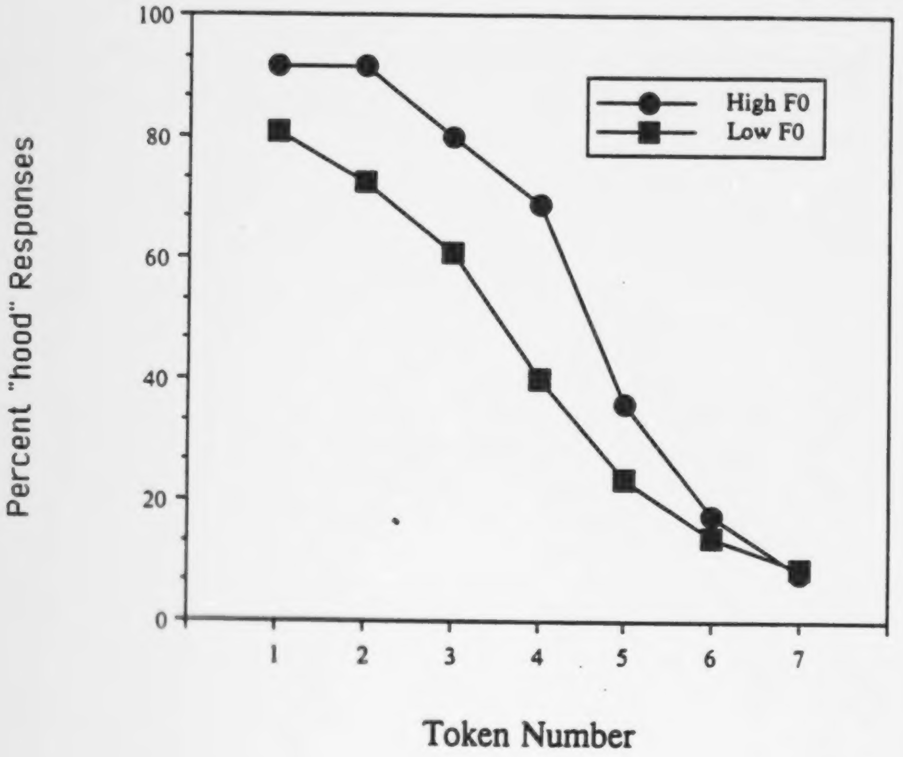


Figure 2. Results of Experiment 2. Identification functions for the high and low F0 items.

that vowels produced by the higher-pitched voice will be identified more often as "hood" than the vowels produced by the lower-pitched voice. Thus, the prediction of the range normalization hypothesis is that the items in the falling intonational context (higher overall pitch range) will be identified as "hood" more often than the items in the rising intonational context.

Method

Subjects. Nine undergraduate psychology students at Indiana University (5 female, 4 male) served as subjects. The subjects had not participated in the previous experiments and were naive as to the purposes of the study. They all reported no history of speech or hearing disorder, were native speakers of American English, and received partial course credit for their participation.

Materials. The "hood"-"hud" vowel continuum which was used in Experiments 1 and 2 was also used in this experiment. The F0 of the test word was, in both intonational contexts, 150 Hz. Two different versions of the phrase "This is hVd" were synthesized. In the rising version, F0 started at 105 Hz and dipped to 95 Hz during "this" and then rose to 150 Hz by the start of the test word. The falling contour started at 185 Hz, rose to 240 Hz during "this", and then fell to 150 Hz by the start of the test word. The intonational contours are shown in Figure A1 in the Appendix. In the pretest these contours were identified as having been produced by the same speaker only 10% of the time. The falling contour sounded like a talker with a high-pitched voice making a statement and the rising contour sounded like a talker with a low-pitched voice asking a question.

Procedure. The procedure was identical to that used in Experiments 1 and 2 except that there was not a manipulation of presentation type. Only items in intonational contexts were presented.

Results and Discussion

A two way repeated measures ANOVA was performed on the identification data. The factors were TOKEN NUMBER and INTONATIONAL CONTEXT. As expected the TOKEN main effect was significant $[F(6,48) = 80.9, p < 0.001]$. There was also a significant main effect for INTONATIONAL CONTEXT $[F(1,6) = 13.15, p < 0.01]$. Subjects identified items in the falling context as "hood" 52.5% of the time while items in the rising context were labeled "hood" only 44.9% of the time. Also, the interaction between the TOKEN and CONTEXT factors was significant $[F(6,48) = 5.10, p < 0.001]$. As is indicated in Figure 3, the context had an influence on how the ambiguous tokens (#3 and #4) were labeled while the perceptual identities of the endpoints remained stable.

Insert Figure 3 about here

The degree of shift in identification was estimated from the average "hood" responses across all of the tokens in the two types of intonational context (the difference between the %"hood" responses to the falling context and the rising context). The degree of shift in identification (7.6% "hood" responses) is smaller than might have been expected from the results of Experment 1 and the pretest. In Experiment 1, the difference between the % "hood" responses to isolated test words and the % "hood" responses to the items in context was 36.7%, and since, in the pretest, the degree of speaker-identity difference between these intonational contexts was roughly equal to the degree of speaker-identity difference between the isolated tokens used in Experiment 1, I would expect that the degree of shift in this experiment would be about the same as the degree of shift associated with a range normalization process in Experiment 1. This difference in results could have two explanations. First, it may be that the measure of speaker-identity used in the pretest was not fine-grained enough to give a sufficiently accurate measure of the perceived difference between talkers (perhaps a multi-valued dimension would give a more accurate measure than the binary dimension which was actually used). Second, it is a possibility that the tokens in isolation were subject to a larger contrast effect than the items in carrier phrases[7]. Whatever the cause of the difference in magnitude between the two experiments, the general results from both support a view of F0 normalization that involves both an auditory F0 normalization process and a range normalization process (which is sensitive to cues about speaker-identity).

The data of Experiment 3 are directly analogous to the data reported by Ladefoged and Broadbent (1957). However, instead of signalling speaker-identity by shifting the formant values of the carrier phrase (as was done by Ladefoged and Broadbent), the identity of the speaker of these phrases was changed by changing the range of F0 in the carrier phrase. Both studies demonstrate a range normalization effect. In one case, speaker-identity information is present in the formant ranges of the carrier phrase, and in the other case, speaker-identity information is present in the F0 range of the precursor phrase.

Conclusion

In Experiment 1, I found evidence that suggests that when differences in perceived speaker characteristics are reduced by the pitch range of a carrier phrase the degree of shift in vowel identification is reduced but not eliminated. This was taken as evidence for both an auditory F0 normalization process (to account for the residue of shift in the vowel identification functions) and a range normalization process (to account for the fact that reduction of speaker-identity differences resulted in a reduction in the shift in the vowel identification

---

[7] In Johnson (1988) I report the results of two experiments which demonstrate the existence of a quite large contrast effect in F0 normalization for tokens presented in isolation.

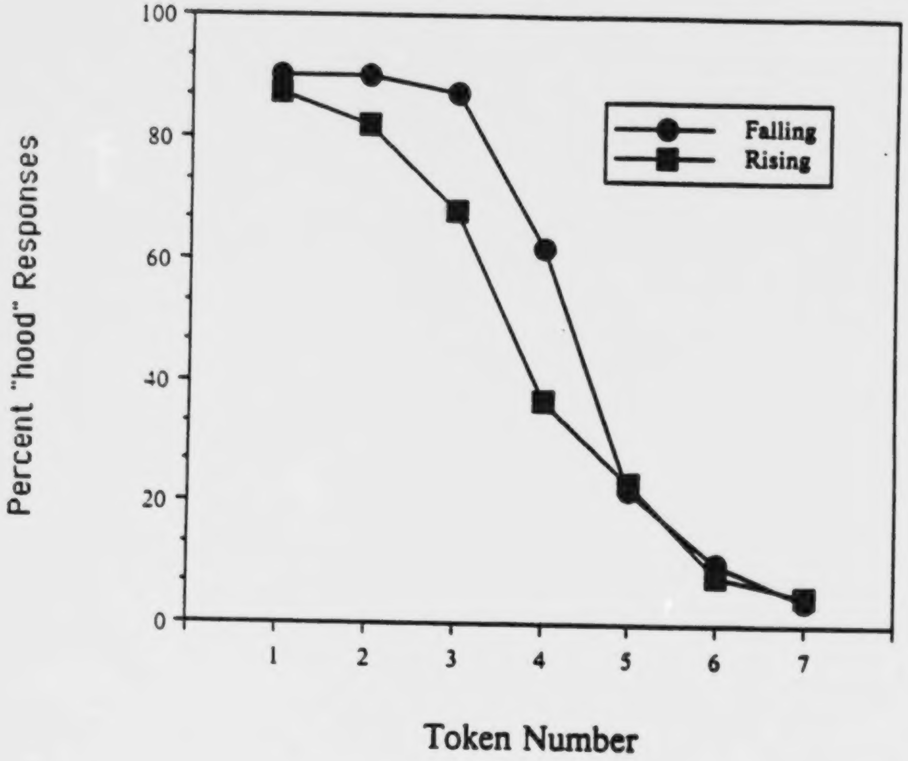


Figure 3. Vowel identification results for the falling and rising intonation contours in Experiment 3.

functions). In Experiment 2, it was shown that if items in carrier phrases and in isolation are matched for degree of speaker-identity difference and F0 difference, the presentation of items in carrier phrases has no impact on vowel identification performance. Experiment 3 demonstrated that speaker-identity differences alone are enough to cause a shift in vowel identification performance. These results were taken as further evidence for the view that F0 normalization involves a range normalization process. The data of the third experiment also suggest that estimated vowel formant ranges can be influenced by pitch range (with no correlated change in formant range of the precursor phrase).

In the introduction, I suggested a model of vowel normalization which includes both an auditory F0 normalization process and a range normalization process. I also suggested that a range estimation process may operate in the perception of isolated syllables in which formant ranges are estimated from vowel-internal phonetic characteristics (particularly F0). The results of the experiments reported here can be interpreted given this model of vowel normalization. They cannot be interpreted, however, in a framework which does not include reference to the perceived phonetic characteristics of the speaker.

Appendix

This appendix describes a pretest which was conducted in order to determine speaker-identity characteristics of the carrier phrases and test words which were used in Experiments 1-3. The pretest was designed to gather speaker-identity data which could then be used in the design and interpretation of the experiments. A pair of tokens (either two intonational contours or two test words in isolation) were presented to subjects who were asked to classify the items in the pair as having been produced by the same speaker or a different speaker. The only acoustic parameter which was manipulated in the tokens was F0. Carrell (1984) found that both F0 and vowel formant range were strong cues for speaker identity. The vowel formants which were used in synthesizing the tokens here were ambiguous between those reported by Peterson and Barney (1952) for men and women. Thus, the perceived identity of the speaker was easily modified by changes of F0.

Method

Subjects. Ten undergraduate students (7 males, 3 females) from Indiana University served as subjects in the experiment. Each subject participated in a single one hour session, and received partial course credit for their participation. They were all native speakers of American English and reported no history of speech or hearing disorders.

Materials. Sixteen versions of the phrase "This is hood" were synthesized. Fifteen of these had falling intonational contours which ended at fifteen different F0 levels (90-160 Hz in 5 Hz steps). The sixteenth version had a rising intonational contour which ended at 150 Hz. The rising contour can be transcribed in Pierrehumbert's (1980) system of intonational transcription as L*HH%, and the falling contours as H*LL%.[8] The fact that the last word in the phrase had steady-state F0 meant that the intontation patterns were somewhat stylized. Formants of [ɩ] in "this" and "is" were between the values for men and women reported by Peterson and Barney (1952) ($F1 = 456$, $F2 = 1740$, $F3 = 2796$ Hz). The word "this" was 415 ms long and was clearly the nuclear syllable in the phrase. F0 contours for all of the intonational contexts are shown in Figure A1.

Insert Figure A1 about here

---

[8] In this transcription system H and L stand for relatively high and low F0 levels, respectively and a starred tone indicates a pitch accent (in this case the accent for the nuclear stress on "this"), the percent symbol indicates that the tone is a "boundary tone" and the unmarked H or L is used to indicate a "phrase accent" - a tone which occurs after a nuclear accent and fills up the time until the phrase boundary. The intonation contours used here are the normal patterns for a simple statement (in the case of the falling contour) or question (rising contour) in American English.

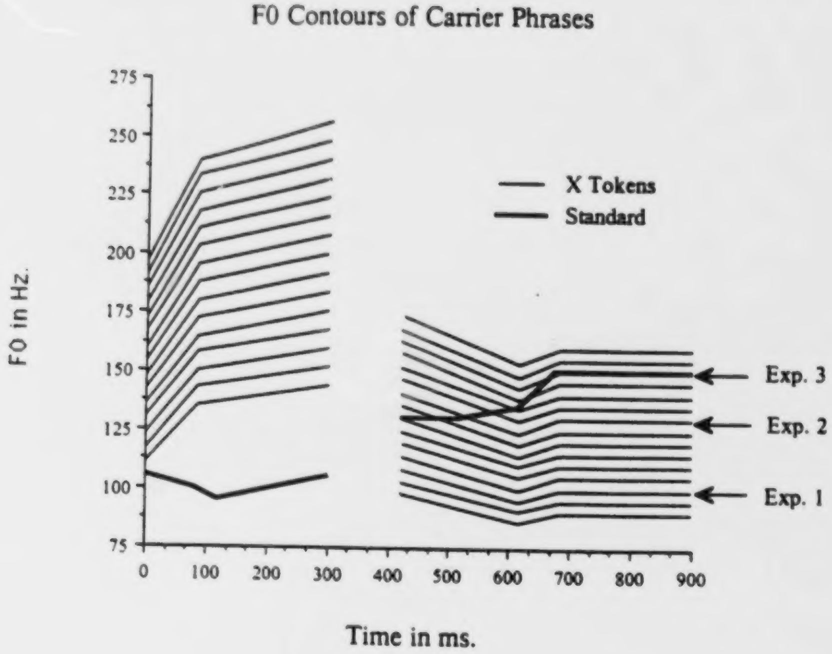


Figure A1. Intonational contours of the tokens used in the pretest.

Fifteen isolated "hood" tokens were synthesized at F0 levels from 90 to 160 Hz in 5 Hz steps. F0 in all of the tokens was steady-state. As with the isolated tokens, the test word in the intonational contours had a steady-state F0. One rising contour was also synthesized. It ended with a steady-state F0 of 150 Hz on the word "hood".

Procedure. The experiment was conducted on-line on a PDP 11/34 computer at Indiana University. Tokens were played at 10 kHz, low pass filtered at 4.8 kHz, and presented binaurally over matched and calibrated TDH-39 headphones at a peak SPL of 85 dB. Subjects were run in groups of up to six at a time, and reaction time as well as response data were collected. Labels for the response buttons used by subjects were displayed on a CRT at eye level and the response to button association was switched for successive blocks to counterbalance for handedness effects in reaction time.

Subjects were asked to judge the relative identity of the speakers of the tokens in a fixed-standard AX task. In this task one token is chosen as a standard against which the other tokens are judged. The subject's task is to decide whether the X token had been produced by the same talker who had produced the A (standard) token. The standard was always presented before the test token. For both the intonational context items and the items in isolation the standard test word had an F0 of 150 Hz. The rising intonation contour was the standard for the intonational context items. The items were presented blocked according to context type (intonational context versus test words in isolation) and the phrases were presented to all subjects before the isolation items.

Results

The speaker classification results are shown in Figure A2. As is clear from the figure there was a large interaction between the test word F0 level and context conditions [F(1,14) = 22.62, $p < 0.001$]. There were some large differences between subjects that indicate that subjects adopted different strategies for the task. For instance, one subject classified all of the isolation items as having been produced by different talkers unless the F0 of the test item was 150 Hz (identical with the standard). For this subject, the task was essentially a pitch matching task at which he was very successful. At the opposite extreme another subject classified all of the isolated tokens as having been produced by the same speaker. For this subject, the task may have been one of identifying vocal tract characteristics or speaker voice source differences which were not present in the tokens. All subjects exhibited the same pattern of results for the context items (with varying degrees of internal consistency and bias).

Insert Figure A2 about here

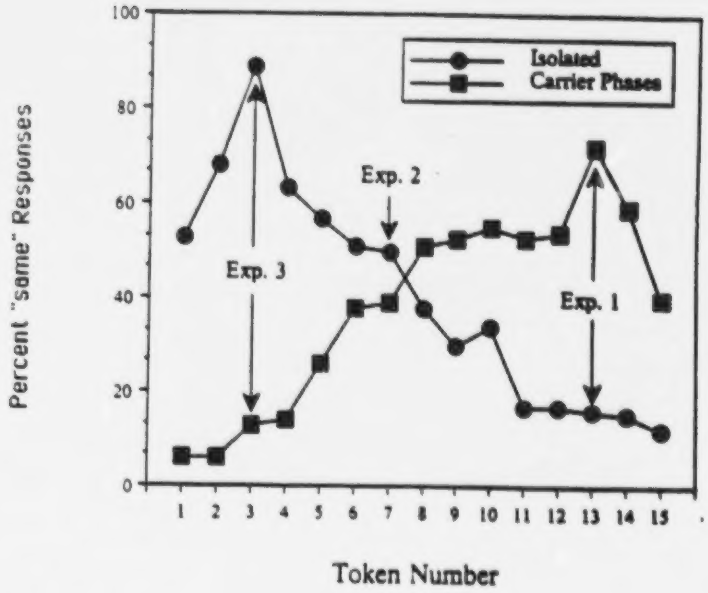


Figure A2. Results of the pretest. For both the items in isolation and the items presented within the intonational contours, the fixed standard test word had F0 of 150 Hz.

The F0 of the test word in the falling intonation tokens which were judged to be most similar to the rising standard contour is quite different from the F0 of the test word in the standard context. The falling context token with test word F0 of 100 Hz was judged to be most similar in speaker characteristics to the standard (75% "same" responses), while the falling context token with test word F0 identical to that of the standard was classified as having been produced by the same talker only 10% of the time. The items which were used in Experiments 1-3 are indicated in Figure A2.

References

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgements. In Fant, G. & Tatham, M., (Eds.), *Auditory analysis and perception of speech.* London: Academic Press.

Assman, P. F. & Summerfield, Q. (in press). Modeling the perception of concurrent vowels: Vowels with the same fundamental frequency. *Journal of the Acoustical Society of America.*

Bladon, R. (1982). Arguments against formants in the auditory representation of speech. In Carlson, R. & Granström, B., editors, *The representation of speech in the peripheral auditory system.* Amsterdam: Elsevier Biomedical.

Carrell, T. (1984). Contributions of fundamental frequency, formant spacing, and glottal waveform to talker identification. *Research on speech perception technical report no. 5.* Bloomington, IN: Indiana University.

Chiba, T. & Kajiyama, M. (1941). *The vowel: its nature and structure.* Tokyo, JP: Tokyo-Kaiseikan.

Chistovich, L., Sheikin, R., & Lublinskaja, V. (1979). 'Centres of Gravity' and spectral peaks as the determinants of vowel quality. In Lindblom, B. & Öhman, S., editors, *Frontiers of speech communication research.* London: Academic Press.

Chistovich, L. A. (1985). Central auditory processing of peripheral vowel spectra. *Journal of the Acoustical Society of America,* **77**, 789–805.

Egan, J. P. (1948). Articulation testing methods. *The Laryngoscope,* **58**, 995–991.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE-AU,* **16**, 78–80.

Hillenbrand, J. & Gayvert, R. (1987). Speaker-independent vowel classification based on fundamental frequency and formant frequency. *Journal of the Acoustical Society of America,* **81**, S93.

Johnson, K. (1988). F0 normalization and adjusting to talker. *Research on speech perception progress report no. 14.* Bloomington, IN: Indiana University.

Ladefoged, P. & Broadbent, D. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America,* **29**, 98–104.

Miller, J. (1987). Auditory-perceptual interpretation of the vowel. *Journal of the Acoustical Society of America,* **81**, S16.

Nearey, T. M. (1987). Static, dynamic and relational properties in vowel perception. *Journal of the Acoustical Society of America*, **81**, S16.

Nearey, T. M. (1978). *Phonetic feature systems for vowels.* Bloomington, IN: IU Linguistics Club.

Papçun, G. (1980). How do different speakers say the same vowels? *UCLA-WPP*, 48.

Peterson, G. & Barney, H. (1952). Control methods used in a study of the identification of vowels. *Journal of the Acoustical Society of America*, **24**, 175–184.

Pisoni, D. B. & Tash, J. (1974). Reaction times to comparisons within and across phonetic categories. *Perception and Psychophysics*, **15**, 285–290.

Plomp, R. (1975). Auditory analysis and timbre perception. In Fant, G. & Tatham, M. A. A., (Eds.), *Auditory analysis and perception of speech.* London: Academic Press.

Potter, R. & Steinberg, J. (1950). Toward the specification of speech. *Journal of the Acoustical Society of America*, **22**, 807–820.

Remez, R., Rubin, P., Nygaard, L., & Howell, W. (1987). Perceptual normalization of vowels produced by sinusoidal voices. *Journal of Experimental Psychology: Human Perception and Performance*, **13**, 40–61.

Sawusch, J. (1977). Peripheral and central processes in selective adaptation of place of articulation in stop consonants. *Journal of the Acoustical Society of America*, **62**, 738–750.

Sawusch, J. & Nusbaum, H. (1983). Auditory and phonetic processes in place perception for stops. *Perception and Psychophysics*, **34**, 560–568.

Sawusch, J., Nusbaum, H., & Schwab, E. (1983). Contextual effects in vowel perception II: Evidence for two processing mechanisms. *Perception and Psychophysics*, **27**, 421–434.

Scharf, B. (1970). Critical bands. In Tobias, J., (Ed.), *Foundations of modern auditory theory, Vol. 1.* New York, NY: Academic Press.

Sussman, H. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, **28**, 12–23.

Syrdal, A. & Gopal, H. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086–1100.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, **69**, 1465–1475.

van Bergem, D., Pols, L., & Koopmans-van Beinum, F. (1988). Perceptual normalization of the vowels of a man and a child. *Speech Communication*, **7**, 1–20.

Verbrugge, R., Strange, W., Shankweiler, D., & Edman, T. (1976). What information enables a listener to map a talker's vowel space? *Journal of the Acoustical Society of America*, **60**, 198–212.

Zwicker, E. (1961). Subdivision of the audible frequency range into critical bands (frequenzgruppen). *Journal of the Acoustical Society of America*, **33**, 248.

RESEARCH ON SPEECH PERCEPTION:
Progress Report No. 14 (1988)
Indiana University

Word Familiarity and Frequency in Visual and Auditory Word Recognition [1]

Cynthia M. Connine [2] and John W. Mullennix

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported, in part, by NIH Training Grant *NS 07134-09* to Indiana University and, in part, by an individual postdoctoral NIMH Traineeship Grant *MH-09357* to the first author. We would like to thank David Pisoni and the Speech Group at Indiana University for helpful discussions and Albrecht Werner Inhoff for comments on the manuscript and the use of his laboratory to conduct Experiment 1.

[2] Cynthia Connine is now with the Department of Psychology, State University of New York at Binghamton.

Abstract

Two experiments are reported which investigate the relationship between printed word frequency and subjective rated familiarity. Subjects performed a lexical decision task on words which were equated in printed frequency but differed in subjective familiarity and on words which were equated in subjective familiarity but differed in printed frequency. Robust effects of familiarity and frequency were found for stimuli presented visually (Experiment 1) and auditorily (Experiment 2). Specifically, high familiar words were responded to faster than low familiar words and high frequent words were responded to faster than low frequent words. The relevance of these results are discussed for lexical access theories and the relative roles of frequency and familiarity.

Word Familiarity and Frequency in Visual and Auditory Word Recognition

Word frequency has been extensively studied in experiments designed to determine its role in lexical access. The importance of word frequency is evidenced by the fact that many models of word recognition have incorporated word frequency effects into their basic architecture (cf. Forster, 1976; Morton, 1969). For example, Forster's search model of lexical access assumes that word frequency effects are at the level of organization of the search process. In this model, lexical access involves searching a subsection of the lexicon based upon partial lexical information (e.g. a syllable or morpheme unit). Each subsection or "bin" is organized by frequency and lexical search is accomplished by a frequency organized search in which high frequency words are checked against the input prior to low frequency words. High frequency words must be searched through prior to low frequency words. One consequence of this architecture is that low frequency words will take longer to access than high frequency words. An alternative model, the logogen model of lexical access (Morton, 1969, 1982), assumes that the lexicon is activated directly and in parallel. In this model, effects of word frequency are placed at the level of the response stage. Responses for high frequency words are evoked faster than responses for low frequency words.

Although word frequency effects have had a major impact on the development of theories of lexical access, two issues have recently been raised with regard to its role. The first issue concerns whether the number of printed occurrences of a word, as measured by Thorndike and Lorge (1944) and Kucera and Francis (1967), is the best assessment of the representation of frequency in the mental lexicon. It has been suggested that word familiarity, a subjective measure based on familiarity ratings, may be a more accurate measure. In a series of experiments, Gernsbacher (1984) demonstrated that word familiarity can account for inconsistent interactions between word frequency and a number of other lexical variables (bigram frequency, concreteness and polysemy) reported in the lexical access literature, particularly for low frequency words. This hypothesis was tested in a series of experiments with low frequency words that varied orthogonally in subjective frequency ratings and either bigram frequency, concreteness or polysemy. She found that familiarity ratings more accurately predicted reaction times than, for example, concreteness, in a lexical decision task. Gernsbacher suggested that experiential familiarity, as determined by subjective ratings, may be a more sensitive measure than printed word frequency in that it may take into account all encounters with a given lexical item. For example, familiarity measures may reflect exposure to words from language production, as well as experience with auditory, visual, and written forms.

The second issue revolves around task specific contributions to word frequency effects (cf. Balota & Chumbley, 1984). Balota and Chumbley have argued that the lexical decision task exaggerates the effects of word frequency. Specifically, these researchers proposed that the lexical decision task involves an evaluation of the stimulus along a familiarity/meaningfulness dimension which occurs after lexical access. They suggested that the familiarity/meaningfulness value for a given word is based on a quick, global assessment of an item's familiarity or meaningfulness. Given a familiarity/meaningfulness value, the item is either quickly categorized as a word, quickly rejected as a nonword or undergoes further, more detailed processing. It was argued that low frequency items were judged to be words more slowly than high frequency items because the low frequency words were less discriminable from nonwords along the familiarity/meaningfulness dimension. This assessment of the lexical decision task contrasts with a great deal of research which has assumed that lexical decision reaction times primarily reflect lexical access processes.

The experiments presented here were designed to investigate some aspects of the relationship between word frequency and word familiarity. One specific purpose of experiments was to determine whether familiarity and frequency reflect separate components of the lexical access process versus an account of familiarity in which subjective ratings are assumed to be a more accurate measure than frequency. In order to address this question, we utilized stimuli which were varied in terms of word frequency and familiarity (see also Nusbaum & Dedina, 1985). Following Nusbaum and Dedina, a number of considerations are relevant in order to manipulate the variables of frequency and familiarity. There are very few, if any, high-frequency/low-familiarity words. That is, words with high printed frequency generally receive high familiarity ratings. In contrast, there are a number of words with low printed frequency which receive a wide range of familiarity ratings. In order to measure familiarity effects for words with similar printed frequency, groups of low-frequency words were chosen that differed in terms of rated familiarity. Similar to Nusbaum and Dedina (1985), we expect that highly familiar words will be responded to faster than low familiar words matched in terms of printed frequency in a lexical decision task. We were also interested in determining the sensitivity of the familiarity variable. Accordingly, we selected stimuli from a broad range of rated familiarity and formed four subgroups of items according to relatively small differences in ratings. In order to determine whether word frequency effects exist when familiarity is held constant, a fifth group of high frequency words were chosen. For this high frequency group, the distribution of familiarity ratings was comparable to the low frequency group with the highest rated familiarity.

A comparison of high and low frequency words equated for familiarity allows a second issue to be addressed. Specifically, if the lexical decision task contains a post lexical meaningfulness/familiarity assessment, are frequency effects still evident when meaningfulness/familiarity (as measured by subjective familiarity ratings) is equated for high and low frequency words? According to Balota and Chumbley's model of lexical decision, lexical items which are equated in terms of familiarity would undergo similar familiarity/meaningfulness assessments. As a consequence, any inflation of the frequency effect due to the post-lexical access component of the lexical decision task would be neutralized. Therefore, any frequency effects found for stimuli matched in familiarity would be attributable to a pre-access component of the lexical decision task.

Experiment 1

Method

Subjects. Fifteen undergraduate students at the University of New Hampshire served as subjects and were paid $5.00 for their participation. No subject reported any hearing or speech disorder.

Materials. Five groups of 35 lexical items were constructed in the following way.[1] Four levels of word familiarity were selected from an online database which contained word frequency (Kucera & Francis, 1967) and subjective familiarity ratings (see Nusbaum, Pisoni, & Davis, 1984 for further details). Familiarity ratings were obtained by Nusbaum et al. from Indiana University undergraduates who were required to rate each word on a scale from 1 to 7. A rating of 1 was assigned to words whose meaning was unknown; a rating of 7 was assigned to words whose meaning was well known. 35 low frequency stimuli (1 to 7 occurences per million) were chosen for four levels of familiarity (3.1-4.0, 4.1-5.0, 5.1-6.0, 6.1-7.0). Words from the low end of the familiarity scale (rated less than 3.0) were not included. Inspection of these words indicated that they were extremely rare and would most likely be treated as nonwords by our subjects. In addition to the four groups of low frequency words, a fifth group of 35 high frequency (greater than 300 occurences per million), high familiarity (6.1-7.0 rated familiarity) words were chosen. In each of the four levels of familiarity, care was taken to choose items such that they were approximately equally distributed across the familiarity range. This was done in order to avoid a skewed distribution of familiarity ratings within any one familiarity level. In addition, the average frequency was approximately equal across each familiarity group.

Pronounceable nonwords were constructed based on a pool of real words (not used in the experiment) from each of the familiarity/frequency ranges previously described. Nonwords were formed from 175 real words by changing one vowel or consonant.

Procedure. Subjects were tested individually in a quiet room. Presentation of the stimuli was controlled by a microcomputer equipped to record reaction times and lexical decisions. All stimuli were presented in uppercase letters and subjects were seated approximately 80 cms from the CRT screen. Each letter extended approximately .4 degrees of visual angle horizontally. Subjects were instructed to indicate whether each item was a word or nonword by pressing an appropriately labeled response key and to make their response quickly and accurately. A fixation was presented on each trial for 500 ms and was followed by a blank screen for 100 ms prior to presentation of the stimulus. The stimulus remained on the CRT screen until the subject made a response. The interval between trials was two seconds and subjects were provided a rest break every 20 trials. Each subject was presented the entire set of 175 real words and 175 nonwords for a total of 350 trials. Stimuli were randomly presented in a different order for each subject. Reaction times were timed from onset of stimulus.

---

[1] The complete set of experimental materials are available from either author.

Results and Discussion

Table 1 displays the mean reaction times and percent correct as a function of familiarity and frequency. Separate ANOVA's were run on the familiarity and frequency data. Consider, first, reaction times for the familiarity variable. As can be seen in Table 1, reaction times decreased as rated familiarity increased. A one- way ANOVA (four levels of familiarity) conducted on the reaction time data revealed a main effect of familiarity ($F(3,42) = 43.85$, $p < .01$). A series of Newman- Keuls comparisons were performed to determine which levels of familarity contributed to the overall main effect of familiarity (familiarity advantage for each comparison indicated in parentheses). Recall that the words in group 6.1-7.0 were judged to be the most familiar in subjective ratings. The largest significant difference in reaction time (160 ms) was found between groups 3.1-4.0 and 6.1- 7.0; also significant was the difference (107 ms) between familiarity groups 4.1-5.0 and 6.1-7.0. Familiarity group 3.1-4.0 was significantly slower (53 ms) than familiarity group 4.1-5.0 and from familiarity group 5.1-6.0 (74 ms). One comparison, the 21 ms difference between familiarity groups 4.1-5.0 and 5.1-6.0, did not reach an acceptable level of significance.

Consider, next, reaction times for high and low frequency words. Low frequency words were significantly slower (55 ms, $t(14) = 118.6$, $p < .01$) than high frequency words. It is important to note that the frequency effect was obtained for words in which familiarity ratings were equated.

Insert Table 1 about here

Next, the effects of familiarity and frequency on lexical decision accuracy are considered. Similar to the results obtained by Gernsbacher (1984) and Nusbaum and Dedina (1985), we found that accuracy decreased as word familiarity decreased. A one-way ANOVA conducted on the accuracy data for the factor of familiarity (four levels of familiarity) revealed a significant main effect ($F(1,14) = 91.7$, $p < .01$). Neuman-Keuls paired comparisons revealed that the familiarity groups differed significantly from one another in the expected direction (familiarity percent correct advantage indicated in parentheses): accuracy in the least familiar group, 3.1-4.0, was less than group 4.1-5.0 (17.7%), group 5.1-6.0 (42.5%) and group 6.1- 7.0 (57.6%). Familiarity group 4.1-5.0 was less accurate than group 5.1-6 (24.8%) and familiarity group 6.1-7.0 (39.9%). Finally, familiarity group 5.1-6.0 was less accurate than

Table 1

Lexical decision reaction times as a function of familiarity and frequency. Percentage correct is indicated in parentheses (Experiment 1).

	Frequency				
	Low				High
Familiarity	3.1-4.0	4.1-5.0	5.1-6.0	6.1-7.0	6.1-7.0
	831	778	757	671	616
	(35)	(52.7)	(77.5)	(92.6)	(92.8)

the most familiar words, group 6.1-7.0, (15.1%).

A t-test conducted on the accuracy scores for the high and low frequency groups revealed that frequency did not significantly influence accuracy ($t < 1$).

Experiment 1 replicated the independent effects of familiarity and frequency (Nusbaum & Dedina, 1985; see also Kreuz, 1987). Rated familiarity was highly predictive of reaction times for groups of words in which word frequency was controlled. That is, low frequency words with high rated familiarity were judged to be words faster and more accurately then the low familiar counterparts. Word familiarity is clearly a factor in the processes involved in the lexical decision task independent from word frequency. In addition, effects of familiarity were demonstrated for a wide range of stimuli grouped into fairly fine grained levels of rated familiarity. The frequency factor also influenced subjects' reaction times. Lexical decision reaction times for high frequency words were faster than for low frequency words for highly familiar words. The major findings will be discussed in further detail in the general discussion.

Finally, it is of methodological interest to note that familiarity ratings obtained in one region of the United States (the Midwest) can be used to predict performance of subjects in a second region (New England). While we consider this a minor methodological point, successful transfer from one region of the country to another suggests that the familiarity ratings obtained by Nusbaum and Dedina may be of general use.

Experiment 2

Experiment 2 was designed to replicate effects of familiarity and frequency found in Experiment 1 in a different domain. Specifically, we were interested in determining if the effects of word familiarity and frequency found in Experiment 1 transferred to the auditory domain. Our interest in the auditory domain stems from a number of considerations. First, while word frequency has been extensively studied in visual word recognition, few studies have extended the results to auditory materials. Given the emphasis on word frequency effects in word recognition theories, it is of interest to determine if frequency effects exist for spoken language. One theory of word recognition, the cohort model, has been developed specifically for auditory word recognition (cf. Marslen-Wilson & Welsh, 1978). The cohort model makes a number of detailed claims concerning auditory word recognition based primarily on the fact, due to the nature of auditory language, it is available in a sequential, left to right order. Of particular interest here is the role of word frequency in the cohort model. While early versions of the cohort model had no explicit assumptions concerning word frequency, recent versions assume that word frequency is represented in the lexicon as relative activation levels (Marslen-Wilson, 1988). High frequency words presumably have higher activation levels than low frequency words. Effects of word frequency have been demonstrated in auditory lexical decision with materials in which a number of factors were carefully controlled such as uniqueness point and recognition point (Marslen-Wilson, 1988). However, these experiments did not control for word familiarity which, as shown in Experiment 1, produces robust effects on lexical decision reaction times.

Secondly, a number of distinct proposals have been offered for effects of word familiarity which incorporate claims concerning modality. Brown (1984) found that rated familiarity was more highly correlated with frequency counts based on spoken language than on written language. This suggests that familiarity ratings may be primarily reflective of experience with spoken language tokens of words. As previously described, Gernsbacher (1984) proposed that rated familiarity captures exposure to a lexical item across all modalities. If so, familiarity effects should be obtained for auditorilly presented materials, that is, regardless of the specific modality of presentation.

Method

Subjects. Forty-five undergraduate students at Indiana University served as subjects for course credit in a one hour session. All subjects were native speakers of English and no subject reported any history of speech or hearing disorders.

Materials. The list of words and nonwords used in Experiment 1 were used for the present experiment. A male speaker recorded each of the items on audiotape in a sound attenuated booth using an Electro Model D054 microphone and a Crown 800 tape recorder. The words were spoken in isolation in citation form. Each item was low pass filtered at 4.8 kHz and converted to digital form via a 12 bit analog to digital converter sampled at 10 kHz. Amplitude levels were digitally equated using a software package designed to modify waveforms.

Procedure. Subjects were tested in groups of six or fewer in a sound treated room. Subjects listened to stimuli over TDH-39 headphones at a level of 80dB. They were instructed to indicate whether each item was a word or nonword by pressing the appropriate button on a labeled response box. The instructions stressed that the subjects were to make their responses quickly and accurately. 500 ms prior to the presentation of each item, a warning light was turned on to indicate to the subject that the next item was to be presented. Presentation of stimuli and collection of data was controlled by a PDP-11/34A computer. Presentation of stimuli was randomized for each group of subjects and all subjects were presented with 175 word and 175 nonword stimuli. Reaction times were measured from the onset of each stimulus.

Results and Discussion

Table 2 displays the mean reaction times and accuracy data as a function of familiarity and frequency for Experiment 2. Inspection of the reaction time data indicate that effects of familiarity and frequency found in Experiment 1 were also observed here. Consider, first, effects of familiarity on reaction times. It appears that as rated familiarity increased, auditory lexical decision times decreased. In order to quantify this observation, a one-way ANOVA was performed on the familiarity data. As expected, the main effect of familiarity was significant (F(1,44) = 64.5, p < .01). Neuman- Keuls comparisons were performed to determine which familiarity groups differed significantly from each other and a number of comparisons were significant. The lowest rated familiarity group (3.1-4.0) was significantly slower than group 4.1-5.0 (95 ms), group 5.1-6.0 (112 ms) and group 6.1-7.0 (235 ms). Similarly familiarity group 4.1-5.0 was significantly slower than group 6.1-7.0 (149 ms) but there was no significant difference from group 5.1-6.0 (17 ms). Finally, familiarity group 5.1-6.0 was significantly faster than group 6.1-7.0 (123 ms).

Insert Table 2 about here

Similar to Experiment 1, the results of Experiment 2 indicated a robust frequency effect for words equated on familiarity. A t-test conducted on the group means for low and high frequency words revealed that the 60 ms advantage for high frequency compared to low frequency words was significant (t(44) = 32.2, p < .01).

The accuracy data are also displayed in Table 2. Accuracy data for the familiarity and frequency variables will be considered in turn. As in Experiment 1, subjects were more accurate in the lexical decision task as rated familiarity increased. A one-way ANOVA revealed that the effect of familiarity on accuracy was significant (F(1,44) = 269.9, p < .01) and Neuman-Keuls comparisons between individual familiarity groups revealed a number of significant differences. Lexical decision accuracy was worse for the lowest rated familiarity group 3.1- 4.0 compared with each of the other familiarity groups (percentage difference in parentheses): group 4.1-5.0 (22.7%), group 5.1-6.0 (44.7%) and group 6.1-7.0 (55.9%). In addition, more accurate performance was obtained for group 5.1-6.0 (22%) and 6.1-7.0 (33.2%) compared to group 4.1- 5.0. Finally, performance was significantly more accurate in the highest rated familiarity group (6.1-7.0) compared with group 5.1-6.0 (11.2%).

Next, consider accuracy for low and high frequency words equated for rated familiarity. Although subjects were slightly more accurate in lexical decisions for high frequency words compared with low frequency words (3.2%), this effect was not significant (t < 1).

The results of Experiment 2 were quite consistent with Experiment 1. Effects of frequency for auditory stimuli were evident in a fashion similar to the visual domain. High frequency

Table 2

Lexical decision reaction times as a function of familiarity and frequency. Percentage correct is indicated in parentheses (Experiment 2).

			Frequency		
		Low			High
Familiarity	3.1-4.0	4.1-5.0	5.1-6.0	6.1-7.0	6.1-7.0
	1328	1233	1216	1093	1033
	(38.7)	(61.4)	(83.4)	(94.6)	(97.8)

words were identified as words faster than low frequency words in auditory lexical decision when the two groups of words were equated in terms of subjective familiarity. Experiment 2 also showed that words with similiar low printed frequency showed robust familiarity effects. The implications for models of auditory (and visual) word recognition theories are discussed in the following section.

General Discussion

To summarize, effects of familiarity were found for a group of low frequency items that varied in terms of subjective familiarity for both auditory (Experiment 1) and visually (Experiment 2) presented words. Highly familiar items were judged to be words faster and more accurately than less familiar items in a lexical decision task. Words equated for printed frequency nevertheless showed effects of familiarity. It appears that subjective familiarity ratings successfully measure a psychologically real component of lexical processing that frequency measures do not. As such, the results of the present experiment contrast with claims (Gernsbacher, 1984) concerning word familiarity in which subjective ratings of familiarity are suggested to be a more accurate estimate of experiential exposure to a lexical item than printed frequency counts.

Effects of frequency were also obtained for a group of highly familiar lexical items in a visual (Experiment 1) and auditory (Experiment 2) lexical decision task. These results have implications for the analysis of the lexical decision task offered by Balota and Chumbley (1984). Balota and Chumbley suggested that an assessment of an item's familiarity/meaningfulness inflated frequency effects in the lexical decision task. If Balota and Chumbley's account of frequency effects in the lexical decision task is essentially correct, then one would expect that frequency effects would be attentuated, if evident at all, under conditions in which familiarity is controlled. As previously described, the high and low frequency items used in Experiments 1 and 2 were equated for familiarity. Each of the items used were rated as being highly familiar yet robust frequency effects were obtained. At least two alternative conclusions are permitted on the basis of the effects of frequency in the experiments presented here. One possiblity is that while Balota and Chumbley may be essentially correct in their assumption of significant post lexical frequency effects, the present results suggest a robust pre-access component of the frequency effect. If this is the case, these results suggest a significant continuing role for frequency effects in theories of lexical access. It is also possible that the nature of the familiarity assessment in the lexical decision task assumed by Balota and Chumbley is not reflected in the subjective familiarity ratings obtained by Nusbaum, Pisoni and Davis.

While the present investigation provided some additional information concerning the relationship between familiarity ratings and frequency effects, a number of related issues remain for future research. First, it is not clear what knowledge subjects used in producing a familiarity rating. One study has attempted to determine which variables enter in to subjective familiarity ratings. Brown and Watson (1987) attempted to assess the independent contributions of a number of factors to familiarity ratings. Among the variables tested, age of acquisition was more highly correlated with familiarity ratings than either written frequency (Kucera & Francis, 1967) or spoken frequency (Brown, 1984). These researchers concluded that the two frequency measures, as well as age of acquisition, constitute independent contributions to familiarity ratings.

A second issue concerns how familiarity may be incorporated in a theory of lexical access. As previously described, a number of lexical access models have been offered with explicit assumptions concerning the way in which frequency influences lexical processing while few models have attempted to account for familiarity effects. One exception is a recently proposed model of lexical access, the resonance model, which explicitly incorporates subjective familiarity (Gordon, 1985). Briefly, Gordon suggests that activation of a word depends upon the subjective familiarity; the activation function for highly familiar words is steeper than low familiar words. This model adequately accounts for lexical decision reaction times for a wide range of rated familiarity.

Finally, recent theoretical accounts of the lexical decision task preclude straightforward interpretations of the familiarity effects as simply influencing the lexical access process. As noted previously, Balota and Chumbley argued that the lexical decision task involves a post-lexical stage in which familiarity/meaningfulness is assessed. Other researchers have attributed decision processes to the lexical decision task (Seidenberg, Waters, Sanders, & Langer, 1984). In these experiments, a number of context effects that were found in a lexical decision task were not found in a naming task. Seidenberg et al. attributed these task specific patterns of facilitation to an additional post-lexical processing component in the lexical decision task.

In light of analyses of the lexical decision task, it is of considerable interest to determine whether independent effects of familiarity and frequency are found for the naming task. Accounts of the lexical decision task which point to post access components in the response (described previously) have also suggested that naming may be a measure which more accurately assesses lexical access processes (compared to post access processes) (but see Balota & Chumbley, 1985 for evidence implicating production factors in the naming task). Independent effects of printed frequency from subjective familiarity in the naming task would have implications for models of visual and auditory word recognition, particularly for those models which assume a pre-lexical effect of frequency. Experiments using the naming task may provide additional insights into the locus of the effects demonstrated in the experiments presented here and we are currently conducting these experiments.

References

Balota, D.A., & Chumbley, J.I. (1984). Are lexical decisions a good measure of lexical access? The role of word-frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Balota, D.A. & Chumbley, J.I. (1985). The locus of word- frequency effects in the pronuciation task: lexical access and/or production? *Journal of Memory and Language*, **24**, 89-106.

Brown, G.D.A. (1984). A frequency count of 190,000 words in the London-Lund Corpus of English Conversation. *Behavior Research Methods, Instruments,and Computers*, **16**, 502-532.

Brown, G.D.A. & Watson, F.L. (1987). First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory and Cognition*, **15**, 208-216.

Forster, K.I. (1976). Accessing the mental lexicon. In R.J. Wales & E.C. T. Walker (Eds.), *New approaches to language mechanisms*. Amsterdam: North-Holland.

Gernsbacher, M.A. (1984). Resolving 20 years of inconsistent interactions between lexical familiarity and orthography, concreteness and polysemy. *Journal of Experimental Psychology: General*, **113**, 256-281.

Gordon, B. (1985). Subjective frequency and the lexical decision latency function: implications for mechanisms of lexical access. *Journal of Memory and Language*, **24**, 631-645.

Kreuz, R.J. (1987). The subjective familiarity of English homophones. *Memory and Cognition*, **15**, 154-168.

Kucera, H., & Francis, W.N. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Marslen-Wilson, W. (1988). Activation, competition and frequency in models of spoken word recognition. Paper presented at Cognitive models of speech processing: Psycholinguistic and computational perspectives. Sperlonga, Italy. May, 1988.

Marslen-Wilson, W. & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Morton, J. (1969). The interaction of information in word recognition. *Psychological Review*, **76**, 165-178.

Morton, J. (1982). Disintegrating the lexicon: An information processing approach. In J. Mehler, S. Franck, E.C.T. Walker, & M. Garrett (Eds.), *Perspectives on mental representations*, (pp. 89-109). Hillsdale, NJ: Erlbaum.

Nusbaum, H., & Dedina, M. (1985). The effects of word frequency and subjective familiarity on visual lexical decisions. *Research on speech perception progress report no. 11.* Bloomington, IN: Indiana University.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10.* Bloomington, IN: Indiana University.

Seidenberg, M.S., Waters, G.S., Sanders, M., & Langer, P. (1984). Pre- and post-lexical loci of contextual effects on word Perception. *Memory and Cognition*, **12**, 315-328.

Thorndike, E.L., & Lorge, I. (1944). *The teacher's word book of 30,000 words.* New York: Teachers College Press, Columbia University.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 14 (1988)
Indiana University

Similarity Neighborhoods of Spoken Two Syllable Words: Retroactive Effects on Multiple Activation [1]

Michael S. Cluff and Paul A. Luce [2]

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported by NIH Training Grant *NS-07134-10* and NIH Research Grant *NS-12179-12* to Indiana University in Bloomington, IN. The authors would like to thank David Pisoni, Jan Charles-Luce, Steve Goldinger, John Logan, and Mike Dedina for their help in all stages of the study.

[2] Paul Luce is now with the Department of Psychology, State University of New York at Buffalo.

Abstract

The perceptual identification of spondees, two-syllable words consisting of two one-syllable words ("comeback", "suitcase"), was examined. The selection of spondees depended on the neighborhood characteristics of the component words, as determined by the neighborhood activation model (Luce & Pisoni, 1987; Luce, 1986). An "easy" syllable was defined as a high frequency word in a sparse neighborhood of low frequency words, while a "hard" syllable was a low frequency word in a high density, high frequency neighborhood. Each spondee fell into one of four categories depending on the neighborhood structure of its component syllables: Easy-Easy, Easy-Hard, Hard-Easy, and Hard-Hard. In the first experiment, the spondees were created by splicing recordings of each component word of the spondee, thus equating them for syllable stress. Additional experiments tested the perceptual identification of naturally produced spondees, spliced nonwords, naturally produced nonwords, and the component words of the spondees. Neighborhood structure was found to have a strong effect on recognition in all experiments. Also, subjects performed more poorly for spondees with an "Easy-Hard" process than with a "Hard-Easy" pattern, indicating a retroactive temporal pattern in word recognition. These results strongly suggest that word recognition involves multiple activation and delayed commitment, thus ensuring accurate and efficient recognition.

Similarity Neighborhoods of Spoken Two Syllable Words: Retroactive Effects on Multiple Activation

The process of spoken word recognition involves the mapping of information in the speech waveform onto one of tens of thousands of items stored in the mental lexicon. Although this process is rapid and efficient, it is now clear that discriminating among multiple lexical items in memory is crucially dependent on the structural relations among the sound patterns of words in memory (Luce & Pisoni, 1987; Marslen-Wilson, 1987). In particular, the speed and ease with which a spoken word is recognized are directly dependent on the number and nature of sound patterns that must be discriminated among in memory.

Luce and Pisoni (1987) have recently proposed a model of spoken word recognition that attempts to account for the effects of the structural organization of sound patterns on spoken word recognition. This model, called the neighborhood activation model (NAM), proposes that stimulus input (i.e., a spoken word) activates a similarity "neighborhood" of phonetically similar words (i.e.,"neighbors") within the mental lexicon. Two factors are involved in characterizing the structure of a similarity neighborhood. The first, *neighborhood density*, refers to the number of phonetically similar words within a particular neighborhood. Neighborhoods with many words are defined as "dense," whereas those with relatively few words are defined as "sparse." The second main variable in characterizing the structure of a neighborhood is the frequency of words within the neighborhood, or *neighborhood frequency*.

According to NAM, word recognition is initially a stimulus-driven, bottom-up procedure. The stimulus input determines the acoustic-phonetic patterns that are activated in memory. The strength of activation of each of these patterns is a function of the degree of similarity between the item in memory and the stimulus input.

Word decision units, activated by the acoustic-phonetic patterns, then monitor the activation level of those acoustic-phonetic patterns. In addition, these units monitor higher-level lexical information, such as word frequency. Once a decision unit surpasses criterion, word recognition occurs. Luce and Pisoni (1987) have proposed that the processes by which word decision units operate may be described by the Neighborhood Probability Rule. This rule simultaneously takes into account stimulus word intelligibility, stimulus word frequency, neighborhood confusability, and neighborhood frequency:

$$p(ID) = \frac{p(STIMULUS\ WORD) * FREQs}{p(STIMULUS\ WORD) * FREQs + \sum_{j=1}^{n}\{p(NEIGHBORj) * FREQj\}} \quad (1)$$

Based on R.D. Luce's (1959) choice rule, the Neighborhood Probability Rule states that the probability of identifying a particular stimulus word equals the frequency-weighted probability of the stimulus word ($p(STIMULUS\ WORD) * FREQs$) divided by the frequency-weighted probability of the stimulus word plus the sum of the frequency-weighted probabilities of its neighbors $(\sum_{j=1}^{n}\{p(NEIGHBORj) * FREQj\})$.

Luce & Pisoni (1987) obtained support for these predictions using a perceptual identification task. The results showed that words with high stimulus word probabilities were identified more accurately than words with low stimulus word probabilities. Also, words with high neighborhood probabilities (i.e. those within a dense, high frequency neighborhood) were identified less accurately than words with low neighborhood probabilities (i.e within a sparse, low frequency neighborhood). These results demonstrate that neighborhood structure is an important factor in the process of spoken word recognition, and that words are recognized in the context of similar words activated in memory.

Previous tests of the neighborhood activation model have focused solely on the recognition of monosyllabic stimulus words. The goal of the present study was to extend the model by investigating the effects of similarity neighborhood structure on the recognition of two syllable words. In particular, the present study examined the interaction of similarity neighborhood structure and syllable position in spondees. Spondees are two syllable words consisting of two monosyllabic words (i.e., "saucepan," "hacksaw"). The investigation of the recognition of spondees was undertaken for a variety of reasons. Because spondees consist of two real monosyllabic words, this enables direct comparison of the recognition of syllables in isolation and in the context of a two syllable word in which their joint influence on each other can be examined. In addition, the fact that both syllables of a spondee constitute real words allows computation of similarity neighborhoods for the syllables using the same algorithms employed in the previous work on monosyllables. Finally, the stress patterns of component syllables in spondees are approximately equal; control for stress differences in experimental stimuli should therefore result in fairly natural sounding items.

The spondees that were selected had initial and final syllables that were designated as "Easy" or "Hard" depending on their neighborhood structures. "Easy" syllables were high frequency words, according to the Kučera and Francis (1967) word count. "Easy" syllables reside in a neighborhood with the following characteristics: (1) few neighbors and (2) these neighbors have a low mean frequency. "Hard" syllables were low frequency monosyllabic words inside a neighborhood characterized by (1) many neighbors, and (2) high mean frequency. By orthogonally combining each syllable type in each position, spondees were constructed that consisted of four patterns of similarity neighborhood structure: (1) initial easy syllable–final easy syllable (Easy-Easy), (2) initial easy syllable–final hard syllable (Easy-Hard), (3) initial hard syllable–final easy syllable (Hard-Easy), and (4) initial hard syllable–final hard syllable (Hard-Hard).

Using these four sets of spondees, a number of hypotheses regarding the role of similarity neighborhoods in two syllable words were tested using a perceptual identification task. The first question addressed by the present study was the degree to which similarity neighborhoods computed for the individual components of the spondees (i.e., syllables) would predict identification performance for the entire spondee. We predicted that the manipulation of similarity neighborhood difficulty for individual syllables in the spondees would have demonstrable and predictable effects on identification performance for the spondee as a whole. In particular, it was predicted that Easy-Easy spondees would produce the highest level of performance, whereas Hard-Hard spondees would produce the lowest level of performance. Both the Hard-Easy and Easy-Hard spondees were predicted to fall in between the two extremes. In short, we expected that the effects of similarity neighborhood difficulty would combine in an orderly fashion to produce the predicted pattern of results.

The hypotheses of more interest, however, concerned the predicted interactions (or lack thereof) of Hard and Easy syllables on identification performance. These hypotheses were aimed at determining the nature of neighborhood activation in two syllable words and the manner in which ambiguities in the speech signal that arise from densely populated neighborhoods are resolved. Three specific hypotheses were examined: (1) the independence hypothesis, (2) the proactive hypothesis, and (3) the retroactive hypothesis.

Under the independence hypothesis, it was predicted that easy syllables would be identified at equivalent levels of accuracy irrespective of whether they occurred with easy or hard syllables in a spondee. Likewise, under this hypothesis, it was predicted that hard syllables would be identified at equivalent levels of performance irrespective of whether they occurred with easy or hard syllables. In short, the independence hypothesis states that the similarity neighborhood difficulty of one syllable in a spondee should have no effect on the identification of the other syllable and that the resolution of a syllable in its neighborhood should be independent of the adjacent syllable in the spondee.

The proactive hypothesis, on the other hand, predicts a differential effect of one syllable in the spondee on the other depending on syllable position. In particular, the proactive hypothesis predicts that easy syllables in initial position should aid in identifying the final syllable, especially if that syllable falls in a dense, high frequency neighborhood. That is, if the first syllable is relatively easy to identify, then the information in the first syllable may be used proactively to assist in identifying the final syllable. This hypothesis predicts an asymmetry between the Easy-Hard and Hard-Easy conditions. In particular, the Easy-Hard condition should produce higher identification performance than the Hard-Easy condition. This prediction is based on the hypothesis that information is used proactively in order to assist in isolating the target syllable from among its neighbors.

The retroactive hypothesis makes just the opposite prediction. Under this hypothesis, it is predicted that information from one syllable is not used in a predictive manner, but instead is used retroactively in resolving ambiguities that arise in the recognition process. In particular, this hypothesis predicts that easy syllables in final position should aid in identifying syllables in initial position, especially when these initial syllables fall in dense, high frequency neighborhoods that make isolation of a single item difficult. This hypothesis therefore also predicts an asymmetry between the Easy-Hard and Hard-Easy conditions. However, under the retroactive hypothesis, it is predicted that the Hard-Easy condition should produce higher identification performance than the Easy-Hard condition. If information from one syllable is primarily used to resolve ambiguities that have arisen in earlier syllables, then an easy second syllable should result in higher overall identification performance for the Hard-Easy spondees compared to the Easy-Hard spondees, in which the second, hard syllable would do little to assist in increasing identification performance.

In short, three hypotheses were examined regarding the effects of similarity neighborhood structure and syllable position on the identification of spoken spondees. The independence hypothesis states that there should be no differential effect of one syllable on another. The proactive hypothesis states that information from an earlier syllable may be used to assist in identifying a later occurring syllable. Thus, initial easy syllables should increase identification for final hard syllables. The retroactive hypothesis, on the other hand, states that information from later occurring syllables may be used to assist in identifying earlier occurring syllables by helping to resolve ambiguities regarding the identity of the previous syllable.

These three hypotheses all concern the manner in which information in two syllable words is used to resolve ambiguities in the speech signal arising from structural characteristics of the neighborhoods of spoken syllables. Two of these hypotheses, the independence hypothesis and the proactive hypothesis, are consistent with strict left-to-right models of spoken word recognition. These hypotheses claim that when ambiguities arise, they can only be resolved by information within the syllable itself or by previously occurring information. The retroactive hypothesis, on the other hand, is not consistent with a strict left-to-right model because, under this hypothesis, decisions regarding the identity of a spoken syllable may be delayed until later information is processed that may aid in discriminating a previous syllable from among its neighbors.

These hypotheses were tested in a series of experiments that examined the identification of two syllable spoken stimuli in white noise. In each of these experiments, two analyses of the subjects' recognition responses were performed. First, accuracy of identifying the spondee as a whole was examined. Second, in order to examine the possibility of differential effects of similarity neighborhood structure and syllable position, identification accuracy of the component syllables of the spondees was also examined.

Experiment 1

Method

Subjects. Thirty-eight subjects participated in partial fulfillment of requirements for an introductory psychology course. All subjects were native English speakers, had no history of speech or hearing disorders at the time of testing, and were able to type.

Materials. One hundred and forty-four spondees were selected through a search of a computerized version of Webster's Pocket Dictionary. Each spondee contained two monosyllabic words having the pattern consonant-vowel-consonant, vowel-consonant, or consonant-vowel. To ensure that subjects had at least encountered the spondee previously, only spondees with a subjective familiarity rating of four or above on a seven point scale were chosen, where four means "familiar with the word but uncertain of its meaning" and seven means "know the word and its meaning" (Nusbaum, Pisoni, & Davis, 1984). For the monosyllables composing the spondees, only those with familiarity ratings of six or above were used. For both the spondees and the monosyllables, familiarity ratings were approximately equal across conditions. The difficulty of a given monosyllabic word was computed based on the neighborhood characteristics of the syllables. A "hard" word or syllable was a low frequency word in a high-density, high frequency neighborhood. An "easy" syllable was defined as a high frequency word in a low-density, low frequency neighborhood. Frequency of words was determined by the Kučera and Francis (1967) word count. Neighborhoods were computed using confusion matrices for all possible initial consonants, vowels, and final consonants produced by the same speaker used to record the spondees (see Luce, 1986). Based on these neighborhood calculations, a spondee could contain the following patterns of syllables: Hard-Hard, Hard-Easy, Easy-Hard, and Hard-Hard. There were 36 spondees of each type, giving a total of 144 spondees.

In addition to manipulating the variables of neighborhood structure and word frequency, care was taken during the stimulus selection procedure to ensure that the transitional probabilities from the first to second syllables within a spondee and from the second to first syllables within a spondee were approximately equal across conditions. Transitional probabilities were computed using Webster's lexicon by determining the number of possible final syllables given the initial syllable and the number of possible initial syllables given the final syllables.

In the first experiment, spondees were created by splicing together separate utterances of the component syllables, thus controlling for the possible effects of syllable stress on recognition. As a result, the spliced utterance "jigsaw" contained the utterances "jig" and "saw." These monosyllables were recorded by a speaker of Midwestern dialect in a sound attenuated booth. Both sets of words were stored digitally and were equated for RMS amplitude.

Spliced spondees were created by splicing together the digital waveforms of the single-syllable words using WAVES, a digital waveform editor (Luce and Carrell, 1981). Closure durations of naturally produced spondees were measured and used as initial values for closure durations of the spliced spondees. If the resulting spondee sounded "unnatural" either due to an overly long or short closure duration, durations between syllables were corrected by ear. Both authors and a trained phonetician listened to and corrected the stimuli until they were considered acceptable.

Procedure. Stimulus presentation was controlled by a PDP-11/34 minicomputer. Stimuli were presented via a 12-bit digital-to-analog converter over matched and calibrated TDH-39 headphones at a +5dB SPL signal-to-noise ratio. To achieve this ratio, the signal was presented at 75dB SPL and the noise at 70dB SPL.

Subjects were tested in a sound-treated room in individual booths equipped with ADM CRT terminals. On each trial, subjects saw the prompt "GET READY FOR NEXT WORD." One second after the prompt, 70dB SPL of white noise was presented. One hundred msec after the onset of the noise, a randomly selected stimulus was presented to subjects. One hundred msec after the offset of the stimulus, the noise was turned off. Immediately following offset of the noise, a prompt appeared on the CRT in front of the subjects. At the prompt, subjects typed in their responses, which could be seen and corrected by the subject, and terminated their responses by pressing the RETURN key. Responses were collected by the PDP 11/34 and stored in data files for later analysis.

Results and Discussion

Separate analyses were performed for whole-word recognition and syllable recognition. In Figure 1, percent correct identification for whole spondees is shown for each syllable structure.

Insert Figure 1 about here

In order to evaluate the overall effect of similarity neighborhoods on spondee identification, a one-way ANOVA was performed on the accuracy scores. Recognition of the spondees in noise was significantly affected by neighborhood structure [$F(3,111) = 51.83$, $p < .05$, $MS_e = .0046$]. Identification performance was most accurate when the spondee consisted of two Easy syllables, and was least accurate when the spondee consisted of two Hard syllables.

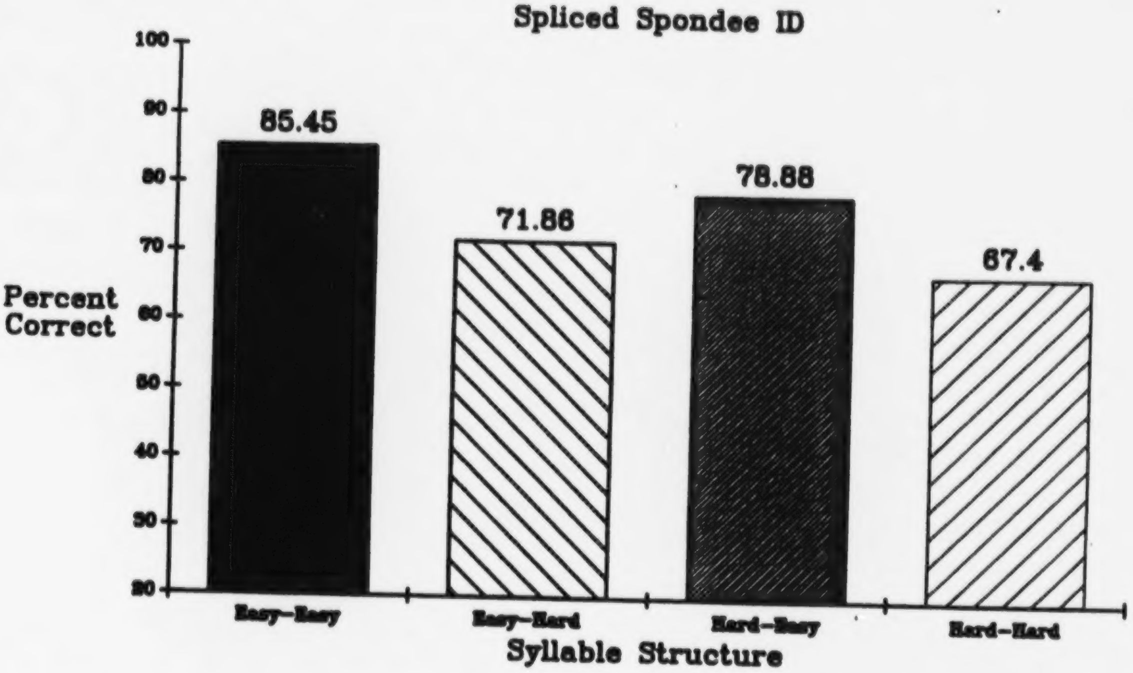


Figure 1. Spliced spondee word identification.

Tukey HSD tests indicated that percent correct identification differed significantly between all syllable structures. Easy-Easy spondees were recognized most accurately, followed by Hard-Easy, Easy-Hard, and Hard-Hard spondees ($p < .05$).

In addition to the whole-word analysis, accuracy scores for the component syllables were also analyzed. The results are shown in Figure 2. Given the neighborhood structure of the component syllables and the hypotheses under test, three variables are of interest in this analysis: the neighborhood characteristic of the syllable (Easy or Hard), the position of the syllable within the spondee (first or second), and the neighborhood characteristics of the other syllable (paired syllable) in the spondee (Easy or Hard). In order to determine the effects of these factors, a 2x2x2 (neighborhood structure X syllable position X neighborhood structure of the paired syllable) ANOVA was performed on the accuracy scores for the individual syllables.

Insert Figure 2 about here

A significant main effect was found for the neighborhood characteristics of the syllable. Easy syllables were recognized more accurately than Hard syllables [$F(1,37) = 165.93$, $p < .05$, $MS_e = .00463$]. In addition, a significant main effect was found for syllable position. Second syllables were identified more accurately than first syllables [$F(1,37) = 17.29$, $p < .05$, $MS_e = .00198$]. A main effect was also found for the syllable paired with the target syllable. Syllables that were paired with Easy syllables were recognized more accurately than those that were paired with Hard syllables, regardless of neighborhood structure [$F(1,33) = 18.34$, $p < .05$, $MS_e = .00419$].

The pattern of whole word results strongly supports the retroactive hypothesis of syllable interaction, due to the more accurate identification performance for Hard-Easy spondees than for Easy-Hard spondees. According to the retroactive hypothesis, ambiguities in the first syllable of a Hard-Easy spondee remain unresolved until the second, easier, syllable is recognized. Because this process is primarily retroactive, the Easy-Hard spondee does not benefit as much from ambiguity resolution, resulting in poorer performance.

The results by syllable not only support the retroactive hypothesis, but also illustrate the interactions between syllables in more detail. The evidence for the retroactive hypothesis can be seen by comparing the identification performance for second syllables with performance for first syllables. Second syllable performance follows a predictable pattern based on the syllables' neighborhood characteristics; Easy second syllables were recognized more accurately than Hard second syllables. First syllable performance, however, does not follow

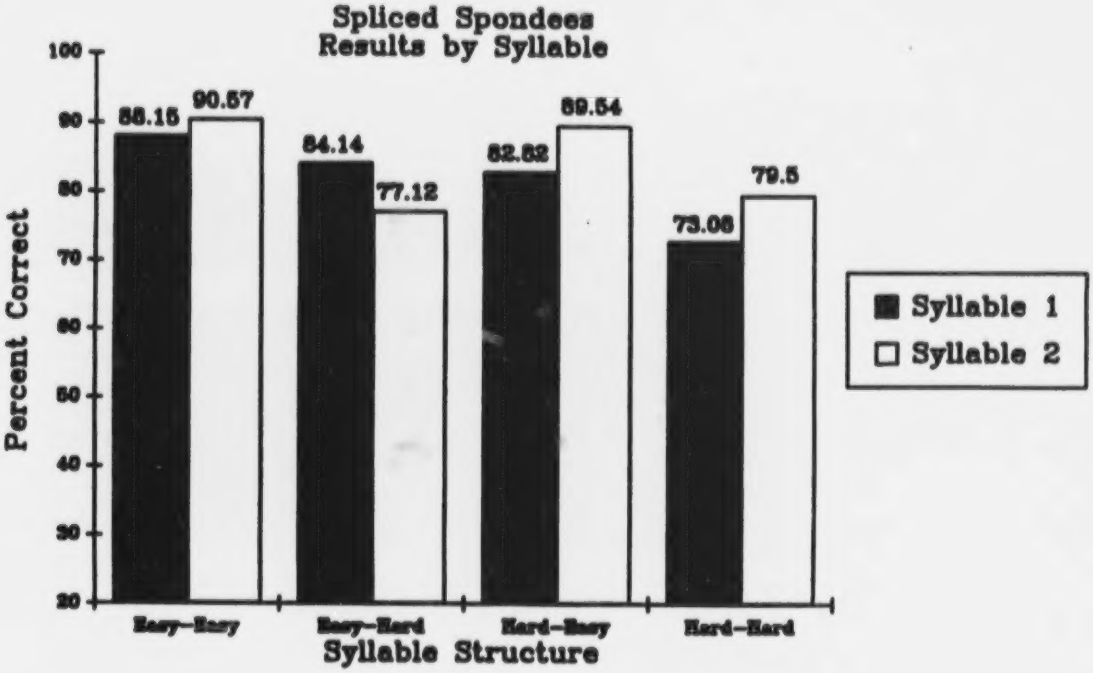


Figure 2. Spliced spondee identification - results by syllable.

this same pattern due to the high performance for the first syllable in Hard-Easy spondees. These Hard syllables were recognized with an accuracy closer to that of Easy syllables, and they were not significantly different from the Easy syllables in Easy-Hard spondees ($p > .05$, $MS_e = .0071$). This suggests that recognition of the Hard syllable in Hard-Easy spondees was enhanced by retroactive resolution from the more easily recognized second syllable.

In order to ensure that the results of Experiment 1 extend to naturally produced spondees, we presented naturally produced versions of the same spondees to subjects in Experiment 2. Though spondees consist of two strong syllables, the first syllable is typically given more stress than the second syllable. Because the spliced spondees were constructed from two separate utterances containing the component syllables, both syllables received approximately equal stress. This might encourage subjects to use an unnatural recognition strategy. In order to test this possibility, Experiment 2 examined identification performance for naturally produced spondees. Given the possible decrease in duration and pitch, one would expect second syllable identification to result in slightly decreased performance, regardless of the pattern of interaction between syllables. Other than decreased second syllable performance, the overall pattern of results for the naturally produced spondees should be similar to the spliced spondee results, if the retroactive hypothesis holds for naturally produced spondees. In other words, final syllable performance should directly reflect neighborhood characteristics, while initial syllable performance should reflect influence from final syllables.

Experiment 2

Method

Subjects. Thirty-four subjects participated to fulfill a course requirement for an introductory psychology course. All subjects were native English speakers, reported no history of speech or hearing disorders at the time of testing, and were able to type. Subjects were drawn from the same source as those used in Experiment 1.

Materials and Procedure. All 144 spondees were recorded in the same manner as the spondees used in Experiment 1, except that the entire spondee was recorded, rather than just the component monosyllables. Thus, stress patterns in the spondees were natural. The stimuli were presented to subjects in the same manner, with the same instructions, and at the same signal-to-noise ratio used in Experiment 1.

Results and Discussion

As in the earlier analyses of the spliced spondees, separate analyses were performed for whole-word recognition and for syllable recognition. In Figure 3, percent correct identification for the entire natural spondee is shown for each syllable structure.

Insert Figure 3 about here

A one-way ANOVA was performed in order to determine the overall effect of similarity neighborhoods on spondee identification. As in Experiment 1, recognition of the natural spondees was significantly affected by neighborhood structure [$F(3,99) = 33.36$, $p < .05$, $MS_e = .00433$]. Tukey HSD tests indicated that Easy-Easy spondees were recognized more accurately than Hard-Hard spondees ($p < .05$). This again supports the prediction that Easy-Easy and Hard-Hard spondees would be the upper and lower end of identification performance, respectively. And, as in the first experiment, Hard-Easy spondees were recognized more often than Easy-Hard spondees ($p < .05$). This result is consistent with the pattern predicted by the retroactive hypothesis.

The results of analysis by syllable for the natural spondees are shown in Figure 4. A 2x2x2 (neighborhood structure X syllable position X neighborhood structure of paired syllable) ANOVA was performed on the accuracy scores for the individual syllables. As before, main effects for syllable type [$F(1,33) = 78.58$, $p < .05$, $MS_e = .00527$] and syllable pairing [$F(1,33) = 5.76$, $p < .05$, $MS_e = .00255$] were found. Again, Easy syllables were recognized more accurately than Hard syllables, and syllables paired with Easy syllables were identified more accurately than syllables paired with Hard syllables. A main effect was also found for syllable position, although first syllables were recognized more accurately than second syllables, which is in the opposite direction from the first experiment [$F(1,33) = 146.00$, $p < .05$, $MS_e = .00189$]. Thus, the prediction that second syllable performance would decrease because of lower stress was confirmed.

Insert Figure 4 about here

The pattern of syllable results also support the retroactive hypothesis. As in Experiment 1, second syllable performance follows a predictable pattern based on the syllables' neigh-

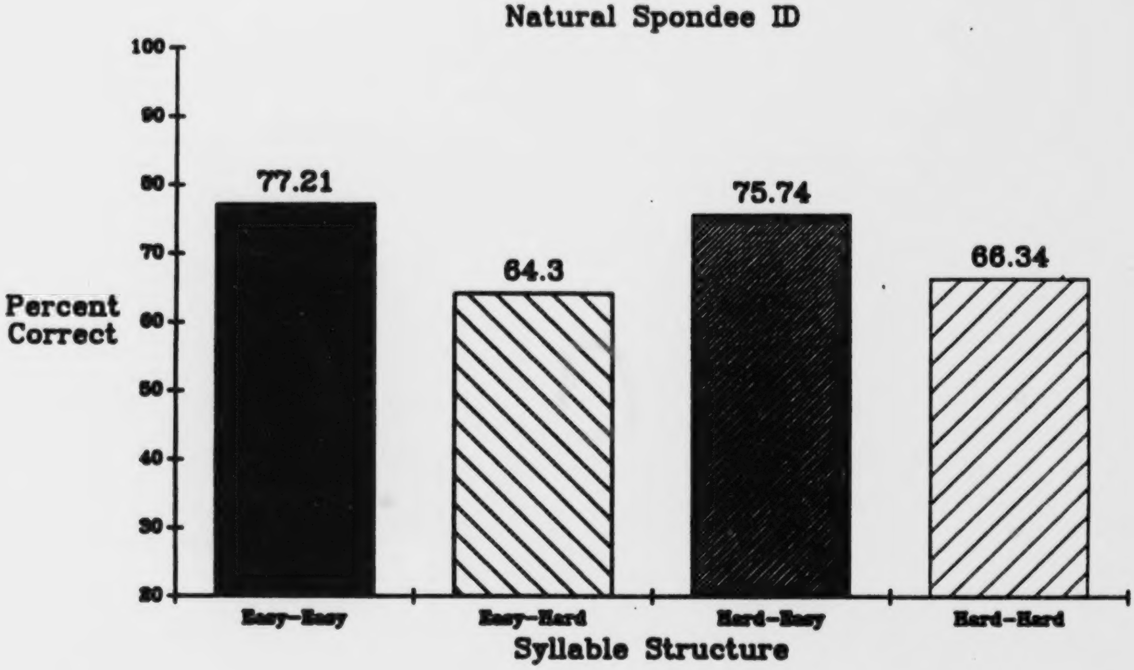


Figure 3. Natural spondee word identification.

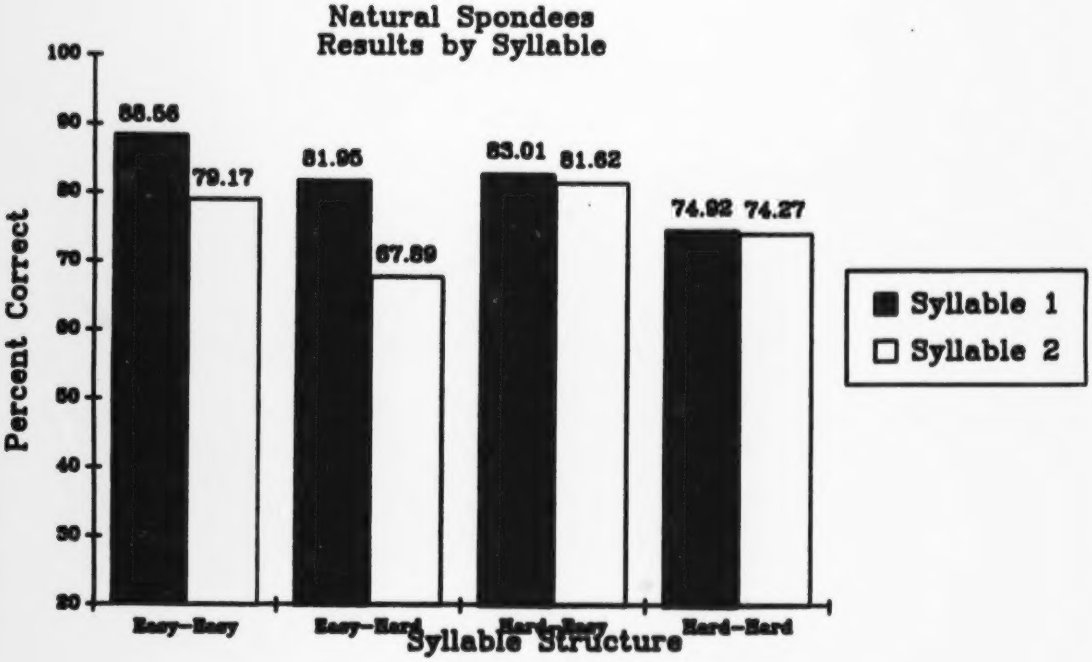


Figure 4. Natural spondees - results by syllable.

borhood characteristics. Easy second syllables were recognized more accurately than Hard second syllables. Performance on first syllables for Hard-Easy spondees was not significantly different from performance on Easy first syllables in Easy-Hard spondees ($p > .05$). This indicates that the Hard first syllable was influenced by the Easy second syllable.

In summary, the results of identification for both spliced spondees and natural spondees support the retroactive hypothesis, even though second syllable performance for natural spondees decreased due to reduced syllable stress. The results of Experiments 1 and 2 indicate that three factors are involved in the resolution of component syllables, and thus recognition for the whole spondee: neighborhood characteristics, syllable stress, and the interaction between syllables. In the following experiment, possible interactions between syllables was eliminated by presenting spliced nonword spondees in a perceptual identification task. Nonword spondees were created by taking all the spondees of one neighborhood structure and randomly re-pairing the second syllables with the first syllables. For example, the Easy-Hard combinations of "fishhook" and "cowhide" were combined to form the nonword, "fishhide."

We predicted that Easy-Easy and Hard-Hard spondees would again produce the extreme levels of performance, with Easy-Hard and Hard-Easy spondees falling in between the two extremes. Because the spondees presented in this experiment were nonwords, no facilatory or inhibitory interaction between syllables should occur. For example, if subjects are presented with the nonword "fishhide" in noise, identification of the Easy syllable ("fish") should not aid in the identification of the Hard syllable ("hide"). As a result, we predicted that performance for Easy-Hard and Hard-Easy spondees would be equivalent.

Experiment 3

Method

Subjects. Thirty-seven subjects participated in partial fulfillment of requirements for an introductory psychology course. All subjects were native English speakers, reported no history of speech or hearing disorders at the time of testing, and were able to type.

Materials and Procedure. Nonword spondees were generated for a given syllable structure by randomly re-combining all first and second syllables within that syllable structure. For example, the syllables of Easy-Hard words "fishhook" and "cowhide" were combined, resulting in the nonword "fishhide." All randomly re-combined spondees that resulted in actual words were re-combined again. As a result of this procedure, the nonword spondees retained the same neighborhood structures as the stimuli in the first two experiments: Easy-Easy,

Easy-Hard, Hard-Easy, and Hard-Hard. The nonword stimuli were generated by splicing together the same recorded monosyllables used in the first experiment. The procedure used was identical to that used in Experiment 1.

Results and Discussion

Figure 5 shows the whole word identification results for nonword spondees. In order to evaluate the overall effect of similarity neighborhoods on nonword spondee identification, a one-way ANOVA was performed on the accuracy scores. Recognition of the nonwords in noise was affected by syllable structure [$F(3,108) = 114.79$, $p < .05$, $MS_e = .00410$]. As predicted, Easy-Easy nonwords were identified most accurately, while Hard-Hard nonwords were identified least accurately. Also as predicted, Tukey HSD results indicated that the Easy-Hard and Hard-Easy nonwords were not statistically different ($p > .05$). The pattern of results suggest that no influence, either proactive or retroactive, was involved in nonword identification.

Insert Figure 5 about here

Results of nonword recognition by syllable are shown in Figure 6. A 2x2x2 (Easy vs. Hard X syllable position X neighborhood characteristics of the paired syllable) ANOVA was performed on the accuracy scores for syllables. A significant main effect for neighborhood characteristics of the syllable was found [$F(1,36) = 644.32$, $p < .05$, $MS_e = .00550$], again showing an advantage for Easy syllables. A main effect for syllable position was also found [$F(1,36) = 4.60$, $p < .05$, $MS_e = .00544$]. Here, first syllables were recognized more accurately than second syllables. In addition, a main effect for pairing was found although it was in the opposite direction from that obtained in the previous experiments [$F(1,36) = 22.79$, $p < .05$, $MS_e = .00271$] (paired with Hard syllable > paired with Easy syllable).

Insert Figure 6 about here

As predicted, the pattern of results for the syllables indicates that no interaction between syllables occurred. Regardless of syllable position, Easy syllables were recognized more accurately than Hard syllables. In the previous experiments, this pattern occurred only for

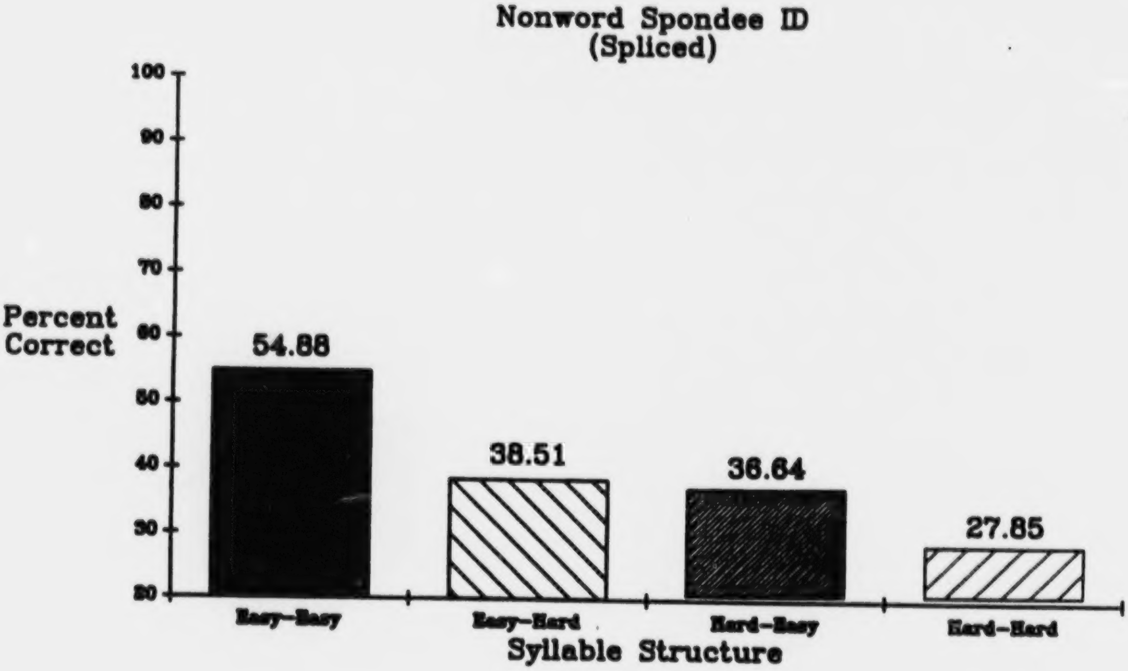


Figure 5. Spliced nonword spondee identification.

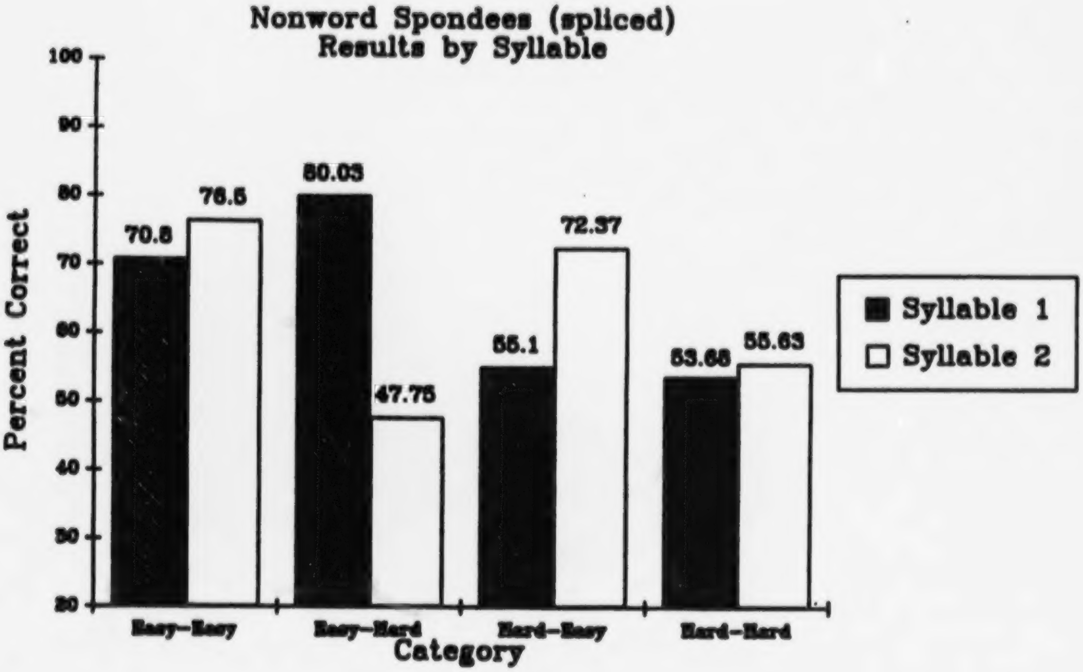


Figure 6. Spliced nonword spondees – results by syllable.

final syllables. Performance for the initial syllables deviated from this pattern previously, as retroactive influence facilitated performance for the Hard syllable in Hard-Easy spondees. In this experiment, however, the Hard syllable in Hard-Easy nonwords was recognized at a performance level comparable to other Hard syllables.

In order to determine if the results of Experiment 3 were due to the peculiarity of nonwords, as well as to assess if the main effects of syllable position and syllable pairing were due to *a priori* differences as a result of the selection process, the component syllables used to create the spliced spondees and the spliced nonwords were presented to subjects in a perceptual identification task. For the current conclusions to hold, we predicted that the monosyllable results should be the same as the syllable results for spliced nonwords.

Experiment 4

Method

Subjects. Forty subjects from a paid summer subject pool participated in the experiment. They were paid $4.00 for their participation. All subjects were native English speakers, reported no history of speech or hearing disorders at the time of testing, and were able to type. All subjects were drawn from the same population of students enrolled at Indiana University.

Materials and Procedure. The monosyllabic recordings used to splice the real and nonword spondees were presented to subjects at the same levels and with the same method as in Experiment 1.

Results and Discussion

Because the syllables were presented separately, analysis was performed for the syllable results alone. These results are shown in Figure 7. A 2x2x2 (Easy-Hard X syllable position X syllable pairing) ANOVA was performed on the accuracy scores for all syllables. Again, a main effect for syllable structure was found, reflecting the advantage for Easy syllables [$F(1,39) = 1378.52$, $p < .05$, $MS_e = .00374$]. A main effect was also found for syllable position. First syllables were identified more accurately than second syllables as in Experiment 3 [$F(1,39) = 12.03$, $p < .05$, $MS_e = .00360$]. The direction of the pairing effect was also identical identical to Experiment 3. Syllables paired with Hard syllables were recognized at a higher level than syllables paired with Easy syllables [$F(1,39) = 32.34$, $p < .05$, $MS_e = .00329$]. Both the syllable main effect and pairing main effect are in the opposite direction as in the first two experiments, suggesting that the current result is due to *a priori* differences

Insert Figure 7 about here

between the syllables.

Recognition performance for individually presented monosyllables was nearly identical to recognition for the component syllables of spliced nonwords. Therefore, the variability of the individual cells in both experiments was due to idiosyncracies within the categories of easy and hard.

General Discussion

We examined the effects of similarity neighborhood structure on the identification of spoken two syllable words in four experiments. In Experiment 1, we examined subjects' accuracy in identifying spondees that were created by digitally splicing together two monosyllabic words. The monosyllabic words that composed the spondees came from either sparsely populated, low frequency neighborhoods (designated as "easy") or densely populated, high frequency neighborhoods (designated as "hard"). The results from Experiment 1 showed reliable effects of similarity neighborhood structure on spondee identification. As predicted, spondees composed of two "easy" monosyllables were identified most accurately and spondees composed of two "hard" syllables were identified least accurately. Furthermore, those spondees composed of one easy and one hard syllable were identified at intermediate levels of performance. Thus, our hypothesis that similarity neighborhood structure would have demonstrable and orderly effects on identification of spondees was confirmed.

Experiment 1 was also aimed at discriminating among three hypotheses concerning possible interactions (or lack thereof) between neighborhood structure (easy vs. hard) and syllable position (initial vs. final). These three hypotheses were: (1) the independence hypothesis, (2) the proactive hypothesis, and (3) the retroactive hypothesis. Under the independence hypothesis, it was predicted that easy and hard syllables would be identified at equivalent levels of accuracy regardless of the type of syllable with which it was paired. This hypothesis predicts that identification of initial syllables will be independent of identification of final syllables, and vice versa.

In contrast, both the proactive and retroactive hypotheses predict interactive effects of one syllable on another. Under the proactive hypothesis, it was predicted that information

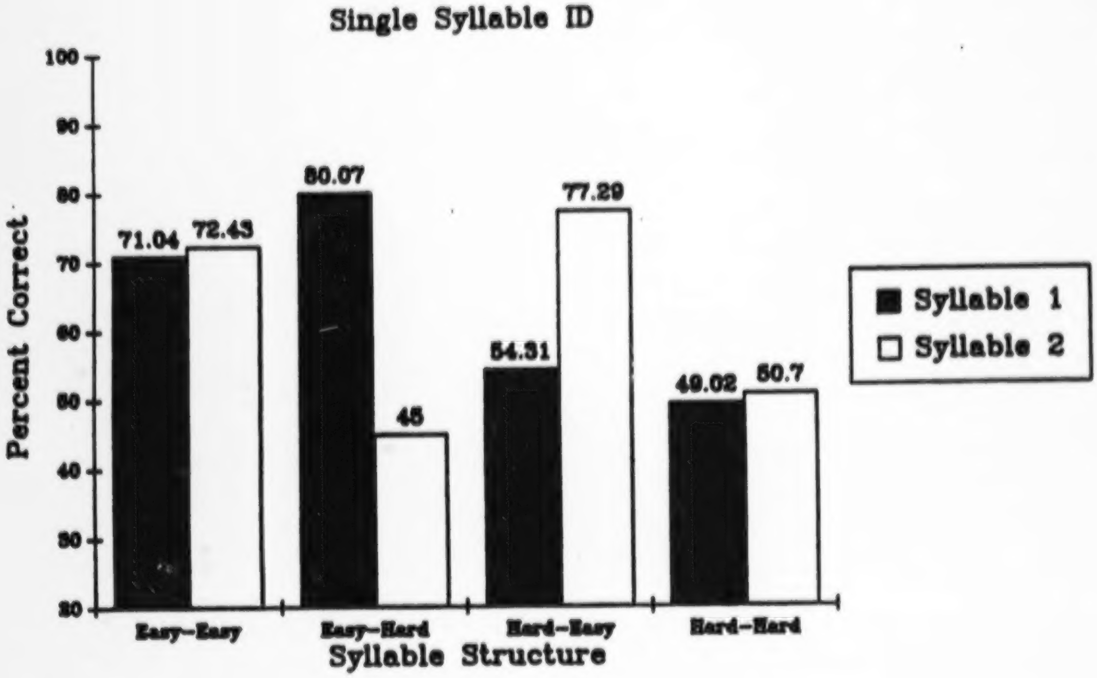


Figure 7. Monosyllable identification.

from previously occurring syllables could be used to aid in identifying later syllables. Specifically, this hypothesis predicts that easy syllables in initial position would aid identification of final syllables, especially when the final syllable falls in a dense, high frequency neighborhood.

Under the retroactive hypothesis, it was predicted that information regarding the possible identity of a syllable would not be used in a predictive manner. Instead, this hypothesis predicts that later occurring information is, in fact, used to resolve previously occurring ambiguities in the signal that arise, in part, from characteristics of similarity neighborhoods. Specifically, this hypothesis predicts that easy syllables in final position would aid identification of initial syllables, especially when the initial syllable falls in a dense, high frequency neighborhood.

The results of Experiment 1 strongly favor the retroactive hypothesis. Identification results for the component syllables in initial and final position revealed differential effects of neighborhood structure on initial and final syllables. In final position, identification accuracy was a direct function of similarity neighborhood structure. Accuracy of identifying syllables in final position was virtually unaffected by the type of initial syllable with which the final syllable was paired. That is, easy final syllables were identified at nearly equal levels of performance regardless of whether the initial syllable with which it was paired was easy or hard. The same pattern of results was obtained with hard final syllables.

In contrast to the results obtained for the final syllables, accuracy of identification for initial syllables was affected by the syllable with which it was paired. In particular, it was found that identification performance for initial easy syllables was reduced when the easy syllable was followed by a hard syllable. Likewise, it was found that identification performance for initial hard syllables was increased when followed by an easy syllable. This is the pattern of results predicted by the hypothesis that later occurring information is used to help resolve previous ambiguities.

In order to ensure that this pattern of results was not due to the artificial manner in which the spondees were constructed, a second experiment using naturally produced spondees was conducted. The results of this experiment showed a pattern of results identical to that obtained in the first experiment, with the exception that accuracy of identification for second syllables dropped overall compared to Experiment 1. This result was due to a slight reduction in the level of stress assigned to second syllables in naturally produced spondees compared to the artificial spondees used in Experiment 1. Nonetheless, the results of Experiment 2 provided support for the hypothesis that information is used retroactively in resolving ambiguities resulting from high levels of competition among words in densely populated, high frequency neighborhoods.

Experiment 3 demonstrated that the pattern of results obtained in the first two experiments was dependent on the lexicality of the spondee. It was hypothesized that if the observed effects were simply due to some unidentified phonetic factor that did not involve accessing the spondee itself, then the pattern of results obtained in Experiments 1 and 2 should also be obtained with nonword spondees. However, if word level factors were responsible for the obtained results, nonword spondees should act according to the independence hypothesis, with no interaction between initial and final syllables. The results from Experiment 3 showed that identification of syllables in nonword spondees is independent of the accompanying syllable, demonstrating that the effects obtained in Experiments 1 and 2 were not due to some lower level phonetic factor but were, in fact, due to retroactive employment of lexical information in ambiguity resolution.

Finally, a fourth experiment was conducted using isolated monosyllabic words to confirm that the identification of component syllables in nonword spondees was in fact independent of effects from the accompanying syllable. The results for the isolated monosyllabic words were virtually identical to the identification of the component syllables of the nonword spondees, further supporting the claim that the results obtained in the first two experiments were due to the interactive use of lexical information in spoken word recognition.

The retroactive effects of lexical information in resolving ambiguities resulting from increased neighborhood competition have important implications for models of spoken word recognition. In particular, the present results suggest that strictly left-to-right models of spoken word recognition, such as cohort theory (Marslen-Wilson & Welsh, 1978) and others (e.g., Cole & Jakimik, 1980) are inadequate. It is clear from the present results that the word recognition system is capable of delaying commitment under conditions of ambiguity (e.g., in the case of a highly competitive neighborhood) and using later occurring information to help resolve this ambiguity. Thus, an appropriate model of spoken word recognition must have some means of maintaining multiple candidates activated in memory and deferring decisions regarding hypotheses about the stimulus input until later occurring disambiguating information is presented. The present results suggest that activation of multiple items or neighbors may result in perceptually ambiguous situations in which the word recognition system may delay decisions.

It should be noted here that the interactive effects of lexical information demonstrated in the present study are primarily retroactive and not proactive or predictive. That is, our results demonstrate a strong asymmetry between the direction of information flow in resolving ambiguity. The finding that the word recognition system prefers to resolve ambiguity based on later occurring information rather than to predict upcoming information may reflect an important design characteristic of the word recognition system and the special signal which it processes. Given our abilities as speakers to produce virtually an infinite number of possible combinations of words in sentences, it is unlikely that precise prediction of upcoming speech input is possible under most normal circumstances in the perception of fluent speech. Thus, the word recognition system may perform optimally by not actively predicting upcoming input; such prediction would likely result in frequent garden-path interpretations. However, if the system avoided active prediction of input and instead maintained multiple items in memory in ambiguous situations, later occurring information could then be employed to resolve previous ambiguities. Indeed, there is considerable evidence that the word recognition system does in fact activate multiple items in memory (Luce, 1986; Marslen-Wilson, 1987). The present study, therefore, serves to elucidate one of the important consequences of the multiple activation of items and how lexical context may be used to choose among these items.

Grosjean and Gee (1987) offer similar proposals regarding the notion of delayed commitment in spoken word recognition. They propose that lexical search is initiated by stressed syllables and that unstressed syllables are recognized by a pattern recognition analysis (see also Cutler & Norris, 1988). Grosjean and Gee state that, in most instances, weak or unstressed syllables will be recognized after lexical search based on the stressed syllables has been completed. Although these proposals are based on issues related to syllable stress, the recognition of stressed vs. unstressed syllables is analogous to the recognition of words in dense and sparse neighborhoods. Indeed, the present results suggest that recognition of unstressed syllables will primarily proceed retroactively after access based on stressed syllables. Evidence for this claim is provided by Grosjean (1985), who demonstrated that reduced words are frequently recognized well after their acoustic offsets once sufficient later, disambiguating information has been processed.

In summary, the present set of studies extend the effects of similarity neighborhood structure on spoken word recognition to two syllable words. The present results demonstrate that the direction of information flow in resolving ambiguities that arise from words in densely populated neighborhoods is markedly asymmetrical. In particular, the process of discriminating and deciding among multiply activated items in memory is primarily affected by subsequent information and not by the generation of predictions or expectancies from prior information in the speech signal. These results demonstrate that the word recognition system operates according to principles of multiple activation and delayed commitment that assure accurate and efficient recognition.

References

Cole, R.A. & Jakimik, J. (1980). A model of speech perception. In R.A. Cole (Ed.), *Perception and production of fluent speech*. Hillsdale, N.J.: Erlbaum.

Cutler, A. & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, 14, 113–121.

Grosjean, F. (1985). The recognition of words after their acoustic offset: Evidence and implications. *Perception and Psychophysics*, **38**, 299–310.

Grosjean, F. & Gee, J. (1987). Prosodic structure and spoken word recognition. *Cognition*, **25**, 135–155.

Kučera, F. & Francis, W. (1967). *Computational analysis of present day American English*. Providence, RI: Brown University Press.

Luce, P. A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6*. Bloomington, IN: Indiana University.

Luce, P.A. & Carrell, T.D. (1981). Creating and editing waveforms using WAVES. *Research on speech perception progress report no. 7*. Bloomington, IN: Indiana University.

Luce, P.A. & Pisoni, D.B. (1987). Neighborhoods of words in the mental lexicon. Submitted manuscript.

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Marslen-Wilson, W.D. (1987). Functional parallelism in spoken word recognition. *Cognition*, **25**, 71–102.

Marslen-Wilson, W. D. & Welsh, A. (1978). Processing interactions during word-recognition in contunuous speech. *Cognitive Psychology*, **10**, 29–63.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10*. Bloomington, IN: Indiana University.

Webster's Seventh Collegiate Dictionary (1967). Los Angeles, CA: Library Reproduction Service.

Similarity Neighborhoods of Spoken Words [1]

Paul A. Luce,[2] David B. Pisoni, and Stephen D. Goldinger

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This is a draft of a paper presented at the workshop on "Cognitive Models of Speech Perception: Psycholinguistic and Computational Perspectives" in Sperlonga, Italy, May, 1988. This research was supported by NIH Research Grant *NS-12179-12* to Indiana University in Bloomington, IN.

[2] Paul Luce is now with the Department of Psychology, State University of New York at Buffalo.

Abstract

A fundamental problem in the study of auditory word recognition concerns the structural relations among the sound patterns of words in memory and the effects these relations have on auditory word recognition. Evidence is reviewed demonstrating that the number and nature of items activated in similarity neighborhoods of words have direct effects on the speed and ease with which spoken words are recognized. This evidence suggests that auditory word recognition is best characterized by two processes: activation of multiple acoustic- phonetic patterns in memory and decision among these competing items. The activation process is assumed to be driven primarily by the similarity of the stimulus word to other words in memory. Following activation, biased decision processes attempt to choose among items activated in a similarity neighborhood based on acoustic-phonetic evidence as well as higher-level lexical information. In particular, it is proposed that stimulus word and neighbor frequencies serve to bias these decision processes. These proposals will be discussed in the framework of the neighborhood activation model, which describes the processes by which words in a similarity neighborhood are activated and decided among.

Similarity Neighborhoods of Spoken Words

Introduction

A fundamental problem in research on spoken word recognition concerns the processes by which stimulus information in the acoustic- phonetic waveform is mapped onto lexical items in memory. Clearly, given the enormous size of the adult mental lexicon, isolating the sound pattern of one word from tens of thousands of others in memory is no trivial problem for the listener, although this process of dicscrimination normally appears to proceed effortlessly and with little error. How, then, can we characterize the human's amazing ability to efficiently isolate the sound pattern of a given word from among the myriad of possible alternatives?

The set of studies reviewed below was aimed at addressing this important question. In particular, these studies investigated the structural organization of the sound patterns of words in memory and the implications of this organization for spoken word recognition. In each of the studies reviewed below, "structure" is defined specifically in terms of similarity relations among the sound patterns of words. Given that one of the primary tasks of the word recognition system involves discrimination among lexical items, the study of the structural organization of words in memory takes on considerable importance, especially if it can be demonstrated that structural relations influence the ease or difficulty of lexical discrimination, and, subsequently, word recognition and lexical access. By the same token, under the assumption that word recognition involves discrimination among competing lexical items, variations in the ease or difficulty of discriminating among items in memory can enlighten us as to the structural organization of the sound patterns of words.

Assuming, then, that structural relations among words should influence auditory word recognition via the process of discrimination, it is important to determine that structural differences among words actually exist. Previous research (Landauer & Streeter, 1973; Luce, 1988a) has indeed demonstrated that words vary substantially not only in the number of words to which they are similar, but also in the frequencies of these similar words. These findings suggest that both structural and frequency relations among words may mediate lexical discrimination. Investigation of the behavioral effects of these relations should help us to understand further not only the process of lexical discrimination, but also the organization of the sound patterns of words in memory.

The classic issue of word frequency takes on an important role in the investigation of the structural organization of the sound patterns of words. Numerous previous studies (Howes, 1957; Newbigging, 1961; Savin, 1963; Soloman & Postman, 1952) have demonstrated that the ease with which spoken words are recognized is monotonically related to experienced frequency, as measured by some objective count of words in the language. However, little work has been devoted to detailing the interaction of word frequency and structural relations among words (see, however, Triesman, 1978a, 1978b). If word frequency influences the perceptibility of the stimulus word, it may likewise affect the degree to which similar words are treated as likely candidates for recognition. Frequency is important, then, in further specifying the relative competition among items that are to be discriminated among.

The experiments summarized below were aimed specifically at examining the effects of the number and nature of words activated in memory on auditory word recognition. In particular, the effects of the structure of similarity neighborhoods were examined. A similarity neighborhood is defined as collection of words that are phonetically similar to a given stimulus word. Similarity neighborhood structure refers to two factors: (1) the number and degree of confusability of the words in the neighborhood, and (2) the frequencies of the stimulus word and its neighbors.

In the experiments discussed below, similarity neighborhood structure was estimated computationally, using a 20,000 word on-line lexicon based on Webster's Pocket Dictionary (Webster's Seventh Collegiate Dictionary, 1967). Each entry in this lexicon contains: (1) an orthographic representation, (2) a phonetic transcription, (3) a frequency count based on the Kucera and Francis (1967) norms, and (4) a subjective familiarity rating on a seven point scale, ranging from "don't know the word " (1) to "know the word and know its meaning" (7) (Nusbaum, Pisoni, & Davis, 1984). Examples of entries in the lexicon are shown in Figure 1.

Insert Figure 1 about here

The general procedure for computing similarity neighborhood structure using the computerized lexicon was as follows: A given phonetic transcription (constituting the stimulus word) was compared to all other transcriptions in the lexicon (which constituted potential neighbors).[1] By comparing the phonetic transcription of the stimulus word with all other phonetic transcriptions in the lexicon, it was possible to determine the extent to which a given stimulus word was similar to other words (i.e., neighborhood density or confusability). In addition, it was also possible to determine the frequency of the neighbors (i.e., neighborhood frequency) as well as the frequency of the stimulus word itself. Thus, three variables relating to the similarity neighborhood structure were of interest: (1) stimulus word frequency, (2) neighborhood density or confusability, and (3) neighborhood frequency.

The precise hypotheses concerning the effects of similarity neighborhood structure that were tested in the experiments reported below were as follows: First, it was proposed that

---

[1] The precise methods by which a neighbor was defined varied as function of the particular experimental paradigm employed and will be discussed in more detail below.

Examples of entries in Webster's lexicon

ORTHOGRAPHY	TRANSCRIPTION	FREQUENCY	RATING
baby	b'e<bi	62	7.00000
bachelor	b'@C-lX	6	7.00000
bacillus	bx-s'I*lxs	2	3.08333
back	b'@k	967	7.00000
bacon	b'e<k|n	10	7.00000
bad	b'@d	142	7.00000
bade	b'@d	1	3.25000

Figure 1. Examples of entries in Webster's lexicon.

stimulus words occurring in highly dense or confusable neighborhoods would be recognized less accurately and less quickly than words occurring in sparse or less confusable neighborhoods. This hypothesis was motivated by the assumption that high density neighborhoods result in a high degree of competition among phonetically similar items activated in memory. Second, it was hypothesized that stimulus words having high frequency neighbors would be identified less accurately and less quickly than those with low frequency neighbors. The effects of neighborhood frequency were predicted under the assumption that frequency affects the degree of competition among items, high frequency neighbors being stronger competitors than low frequency neighbors. Finally, it was predicted that high frequency words would, in general, be identified more easily than low frequency words. However, it was further predicted that effects of stimulus word frequency would be mediated by the structure of the similarity neighborhoods in which the stimulus word resides in the mental lexicon.

Four experiments examining the effects of similarity neighborhood structure on auditory word recognition are reviewed: (1) perceptual identification of words in noise, (2) auditory lexical decision to nonwords, (3) auditory word naming, and (4) primed perceptual identification. Each of these experiments demonstrate that similarity neighborhood structure has robust effects on the speed and accuracy of auditory word recognition. Furthermore, these experiments provide support for a model of spoken word recognition, the neighborhood activation model, that characterizes the processes of spoken word recognition in terms of neighborhood activation and frequency-biased decision.

Evidence from Perceptual Identification

Luce (1986) has developed a means of quantifying the effects of similarity neighborhood structure on the perceptual identification of spoken words called the neighborhood probability rule. This rule represents a single expression that simultaneously takes into account stimulus word intelligibility, stimulus word frequency, neighborhood confusability, and neighborhood frequency. Based on Luce's (1959) general biased choice rule, this rule has the form:

$$p(\mathrm{ID}) = \frac{p(\text{stimulus word}) * \text{freqs}}{p(\text{stimulus word}) * \text{freqs} + \quad \{p(\text{neighborj}) * \text{freqj}\}}$$

The neighborhood probability rule states that the probability of correct identification of the stimulus word is equal to the frequency- weighted probability of the stimulus word (p(stimulus word) * freqs) divided by the frequency-weighted probability of the stimulus word plus the sum of the frequency-weighted probabilities of the neighbors ({ p(neighborj) * freqj}). In general, this rule states the probability of choosing the stimulus word from among its neighbors. In addition, the probabilities of the stimulus word and its neighbors are weighted by their frequencies of occurrence (taken from the Kucera and Francis, 1967, norms). These frequency weights serve to increase, to greater or lesser degrees, the probabilities of the stimulus word and its neighbors, depending on their objective rates of occurrence.

In order to obtain independent estimates of the probabilities of the stimulus word and its neighbors, the probabilities were estimated from confusion matrices for all possible initial consonants, vowels, and final consonants (a complete description of the rule and how it was computed can be found in Luce, 1986). For example, to determine the probability of the stimulus word /kaet/, the separate probabilities for /k/ given /k/ (p(k—k)), /ae/ given /ae/ (p(ae--ae)), and /t/ given /t/ (p(t—t)) were obtained from confusion matrices for initial consonants, vowels, and final consonants, respectively. These independent probabilities were then multiplied to give an estimate of the stimulus word probability. Note that this probability was determined independently from confusion matrices for initial consonants, vowels, and final consonants.

Computation of the probabilities of the neighbors was carried out in an analogous manner. For example, to determine the probability that /kId/ is a neighbor of the stimulus word /kaet/, the confusion matrices were once again consulted. In the case of computing neighbor probabilities, however, the conditional probability of /kId/ given /kaet/ was computed by finding the component probabilities /k/ given /k/ (p(k—k)), /I/ given /ae/ (p(I—ae)), and /d/ given /t/ (p(d—t)). These component probabilities were then mutliplicatively combined to render an estimate of the neighbor probability.

Using this method of computing stimulus word and neighbor probabilities, the neighborhood probability rule can be computed using the computerized lexicon. Basically, this method involves comparing a given stimulus word to potential neighbors in the lexicon and computing the respective probabilities. These probabilities can then be weighted by the frequencies of occurrence of the words to which they correspond and entered into the rule to generate an estimate of the probability of identifying a stimulus word given its frequency, the confusability of its neighborhood, and the frequencies of its neighbors.

A number of properties of the neighborhood probability rule are worthy of mention. First, the intelligibility of the phonemes of the stimulus word itself will determine, in part, the role of the neighbors in determining the predicted probability of identification. Stimulus words with high phoneme probabilities (i.e., words with highly intelligible phonemes) will tend to have neighbors with low phoneme probabilities, owing to the fact that all probabilities in the confusion matrices are conditional. Likewise, stimulus words with low phoneme probabilities (i.e., those with less intelligible phonemes) will tend to have neighbors with relatively higher phoneme probabilities. However, the output of the neighborhood probability rule is not a direct function of the stimulus word probability. Instead, the output of the rule is dependent on the existence of lexical items that contain phonemes that are confusable with the phonemes of the stimulus word. For example, a stimulus word may contain highly confusable phonemes. However, if there are few actual lexical items (i.e., neighbors) that contain phonemes confusable with those of the stimulus word, the sum of the neighbor word probabilities will be low. The resulting output of the neighborhood probability rule will therefore be relatively high. Likewise, if the phonemes of the stimulus word are highly intelligible, but there are a large number of neighbors that contain phonemes that are confusable with the stimulus word, the probability of identification will be reduced. In short, the output of the neighborhood probability rule is contingent on both the intelligibility of the stimulus word and the number of neighbors that contain phonemes that are confusable with those contained in the stimulus word. Thus, intelligibility of the stimulus word, confusability of the neighbors, and the nature of lexical items all act in concert to determine the predicted probability of identification.

Note also that the frequencies of the stimulus word and the neighbors will serve to amplify to a greater or lesser degree the probabilities of the stimulus word and its neighbors. In the long run, high frequency stimulus words are predicted to produce higher levels of performance. Likewise, high frequency neighbors are predicted to reduce identification performance. However, because frequency in this rule is expressed in terms of the relation of frequency of the stimulus word to the frequencies of its neighbors, the absolute frequency of the stimulus word may have differential effects on predicted identification performance depending on the frequencies of the stimulus word's neighbors. For example, given two stimulus words of equal frequency, the stimulus word with neighbors of lower frequency will produce a higher predicted probability of identification than the stimulus word with neighbors of higher frequency. Simply put, this rule predicts that neighborhood structure will play a role in determining predicted identification performance in terms of the combined effects of the number and nature of the neighbors, the frequencies of the neighbors, the intelligibility of the stimulus word, and the frequency of the stimulus word.

Luce (1988b; see also Luce, 1986) tested the predictions of the neighborhood probability by rule selecting four sets of high frequency words and four sets of low frequency words varying on stimulus word probability and frequency-weighted neighborhood probability. Two levels of stimulus word probability (high and low) were orthogonally combined with two levels of frequency-weighted neighborhood probability (high and low) for a set of high frequency words and a set of low frequency words, rendering eight cells in total. Using the methods of determining stimulus word probability and neighborhood confusability described above, 50 consonant-vowel-consonant words [2] were assigned to each of the eight cells (2 levels of stimulus word frequency X 2 levels of stimulus word probability X 2 levels of frequency-weighted neighborhood probability). These 400 words were randomized, mixed with white noise at a signal-to-noise ratio of +5 dB, and presented to subjects for identification.

The results of this experiment are presented in Figure 2. Results for the high frequency words are shown in the top panel; results for the low frequency words in the bottom panel. Solid lines indicate words with high stimulus word probabilities; dotted lines indicate words with low stimulus word probabilities. Frequency- weighted neighborhood probability is represented on the x axes; percent correct identification is represented on the y axes.

---

[2] Each of the words in this study as well as the following studies were rated as highly familiar (5.5 or above) by subjects on a seven point scale (see Nusbaum, Pisoni, & Davis, 1984). This constraint on stimulus selection was imposed to ensure that the subjects knew the words to be presented.

Insert Figure 2 about here

The results of this experiment clearly support the predictions of the neighborhood probability rule. First, words with high stimulus word probabilities were identified consistently more accurately than words with low stimulus word probabilities. More interestingly, however, was the finding that words with high frequency-weighted neighborhood probabilities were identified less accurately than words with low frequency-weighted neighborhood probabilities. That is, words occurring in neighborhoods densely populated by high frequency words were identified less accurately than words occurring in neighborhoods sparsely populated by low frequency words.

Not surprisingly, high frequency words were, on the average, identified more accurately than low frequency words. However, of interest is the finding that high frequency words were not always identified at higher levels of accuracy. In particular, high frequency words residing in dense, high-frequency neighborhoods were identified less accurately than low frequency words residing in sparse, low frequency neighborhoods, as predicted by the neighborhood probability rule.

These results support the hypothesis that accuracy of identifying spoken words in noise is crucially dependent on similarity neighborhood structure. The results demonstrate that increased neighborhood competition reduces accuracy of identification, and that this competition is influenced by both the number of possible neighbors and their frequencies of occurrence. Furthermore, the present results demonstrate that effects of stimulus word frequency are mediated by the structure of the similarity neighborhood. In short, these results provide support for the hypothesis that words are recognized in the context of similar words activated in memory.

Evidence from Auditory Lexical Decision: Nonwords

Although the previous results strongly implicate the role of similarity neighborhood structure in spoken word recognition, the findings are restricted to identification of words degraded by white noise. In order to test the generality of these effects to stimuli that are not degraded, Luce (1986) performed an auditory lexical decision task using specially constructed nonword stimuli varying in neighborhood density and frequency. The choice of examining decisions to nonwords was motivated by a study by Marslen-Wilson (1987) that failed to find

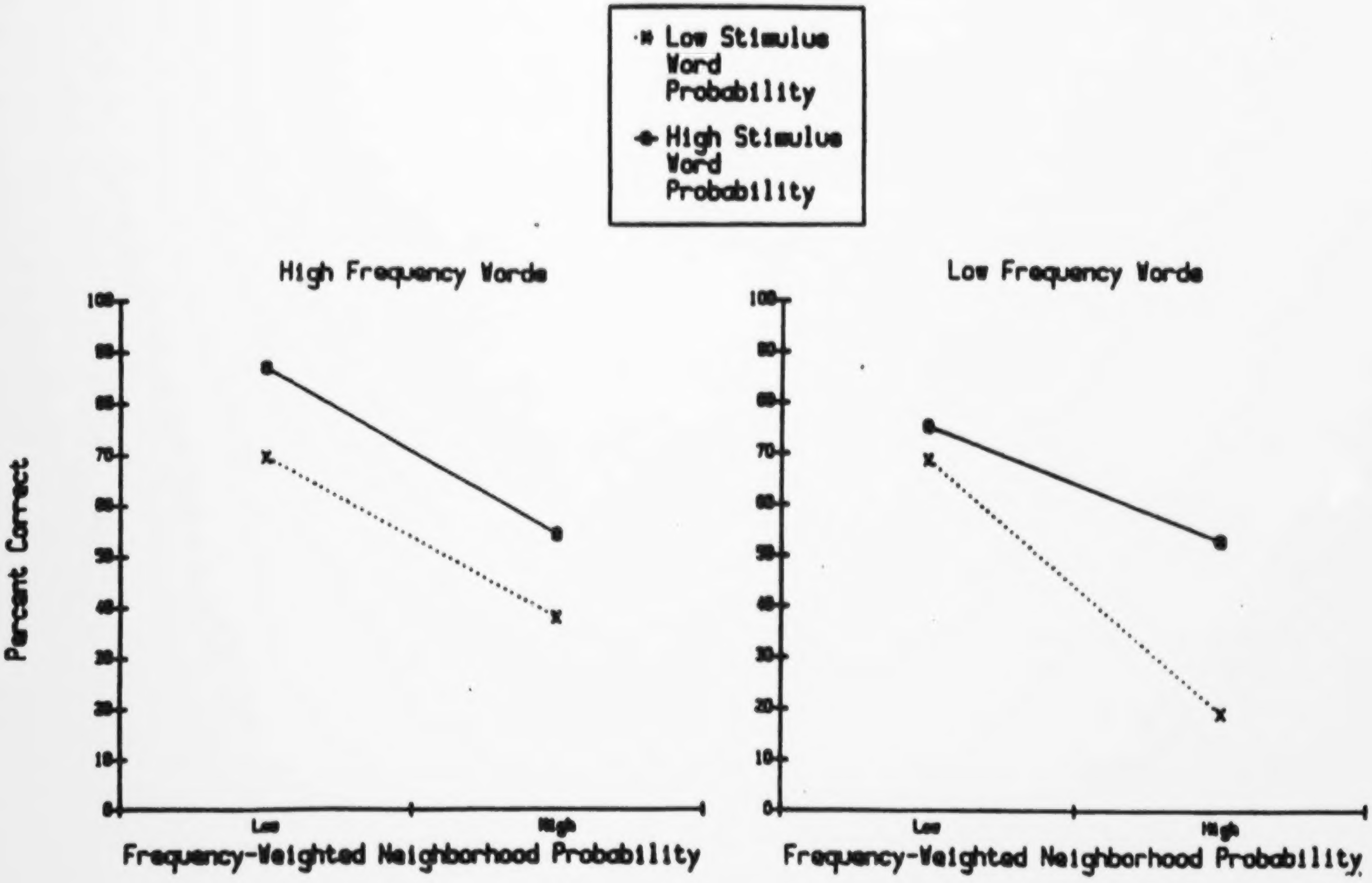


Figure 2. Perceptual identification results. Results for the high frequency words are shown in the top panel; results for the low frequency words are shown in the bottom panel.

set-size (i.e., neighborhood density) effects on reaction times to nonwords in a lexical decision task. Having obtained such effects for perceptual identification, it is of crucial importance to determine if such effects can also be obtained in a task in which the stimuli are not made purposefully difficult to perceive through stimulus degradation manipulations.

Luce (1986) generated a set consonant-vowel-consonant nonwords modeled on the statistical properties of actual words occurring in Webster's Pocket Lexicon. Each of these nonwords had admissible initial consonant-vowel (CV) sequences, vowel-final consonant (VC) sequences, and initial consonant-final consonant (C—C) sequences. In addition, given this manner of stimulus construction, each of the nonwords diverged from real words at the final phoneme. From this set of nonwords, 400 stimuli were selected that fell into one of four cells: (1) high neighborhood density-high neighborhood frequency, (2) high neighborhood density-low neighborhood frequency, (3) low neighborhood density-high neighborhood frequency, and (4) low neighborhood density-low neighborhood frequency.

Neighborhood density and neighborhood frequency were computed as follows: Each nonword was compared to each word in Webster's lexicon. A neighbor was defined as any word that could be converted to the nonword under analysis by a one phoneme addition, substitution, or deletion in any position. For each nonword, the number of neighbors, as well as the frequencies of these neighbors, was tallied and used to assign the nonwords to one of the four cells in the design. These 400 nonwords were mixed with 400 words and presented for word-nonword decisions. Reaction times were measured from the onset of the nonword stimuli to the button-press response. Reaction times for correct nonword judgments are shown in Figure 3.

Insert Figure 3 about here

As can be seen from Figure 3, both neighborhood density and neighborhood frequency had marked effects on nonword judgment times. Nonwords with many word neighbors were responded to significantly more slowly than nonwords with few word neighbors. In addition, nonwords with high frequency neighbors were responded to more slowly than nonwords with low frequency neighbors. Thus, demonstrable effects for both neighborhood density and neighborhood frequency on reaction times to initiate nonword judgments were observed. These results demonstrate that similarity neighborhood effects are not contingent on presentation of degraded stimuli and have marked effects on processing times as well as accuracy of identification.

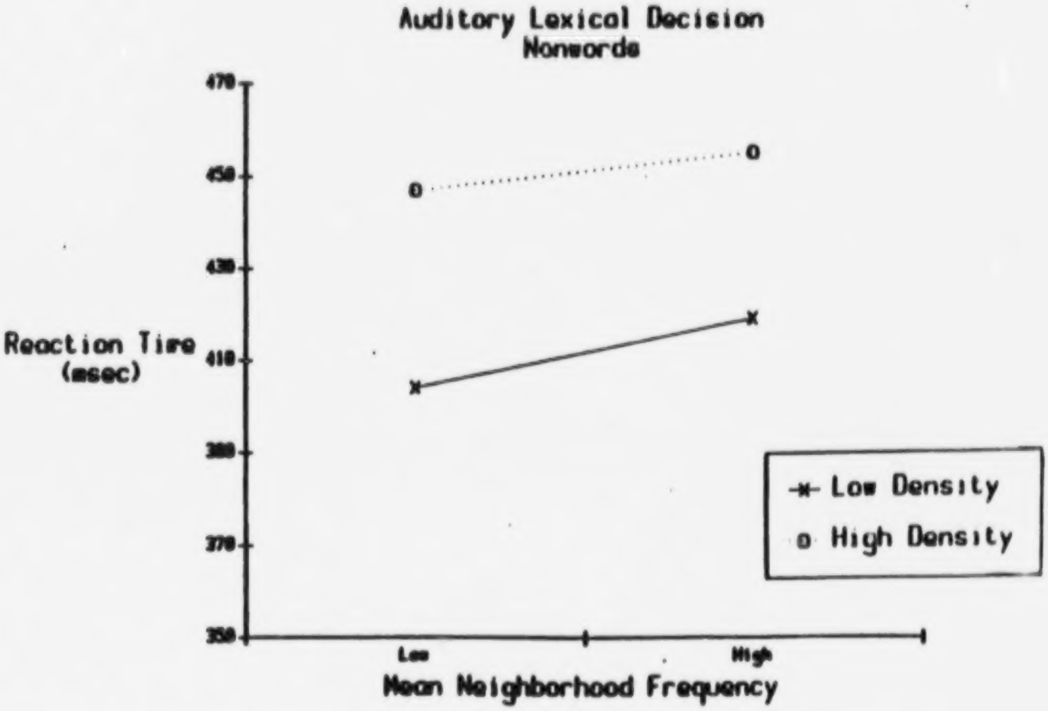


Figure 3. Mean response latencies for nonwords using an auditory lexical decision task.

Evidence from Auditory Word Naming

To further examine the effects of similarity neighborhood structure on spoken word recognition, Luce (1986) performed an auditory word naming experiment using word stimuli varying in frequency, neighborhood density, and neighborhood frequency. In this task, subjects were presented with a spoken word that they were required to repeat as quickly as possible.

The motivation for using the naming paradigm came from recent findings in the visual word recognition literature on the role of word frequency in naming. Balota and Chumbley (1984) have presented evidence demonstrating that word frequency effects are severely reduced in the visual word naming task. Paap, McDonald, Schvaneveldt, and Noel (1986) have further argued that the visual word naming task circumvents lexical access and thus circumvents access to frequency information.

These findings suggest an interesting means of dissociating effects of pattern similarity (e.g., density) and frequency, and thus possibly relegating effects of frequency and multiple activation of neighbors to different levels of processing. In particular, the naming task was employed to determine if frequency effects (both stimulus word frequency and neighborhood frequency) lie at a later, decision stage of processing, whereas multiple activation of neighbors lies at some earlier stage. If the visual and auditory naming tasks are sufficiently similar, it may be hypothesized that frequency effects will be circumvented in the auditory naming task. However, activation of neighbors should not be affected by the task, given that multiple patterns must still be activated and decided upon in order to identify the stimulus word.

In order to test this hypothesis, 400 consonant-vowel-consonant words were were assigned to eight cells. These cells were constructed by orthogonally varying two levels of word frequency (high and low), two levels of neighborhood density (high and low), and two levels of neighborhood frequency (high and low). The method for determining similarity neighborhood structure was identical to that used in the previous lexical decision experiment.

Reaction times for naming the 400 stimuli were measured from the onset of the stimulus to the naming response. Results for the naming task are shown in Figure 4. Because no significant interactions involving the three dependent variables of word frequency, neighborhood density, or neighborhood frequency were observed, only the mean reactions times for each level of the three variables are shown.

Insert Figure 4 about here

Only the effect of neighborhood density was significant. Words in high density neighborhoods were named more slowly than words in low density neighborhoods. No effects of

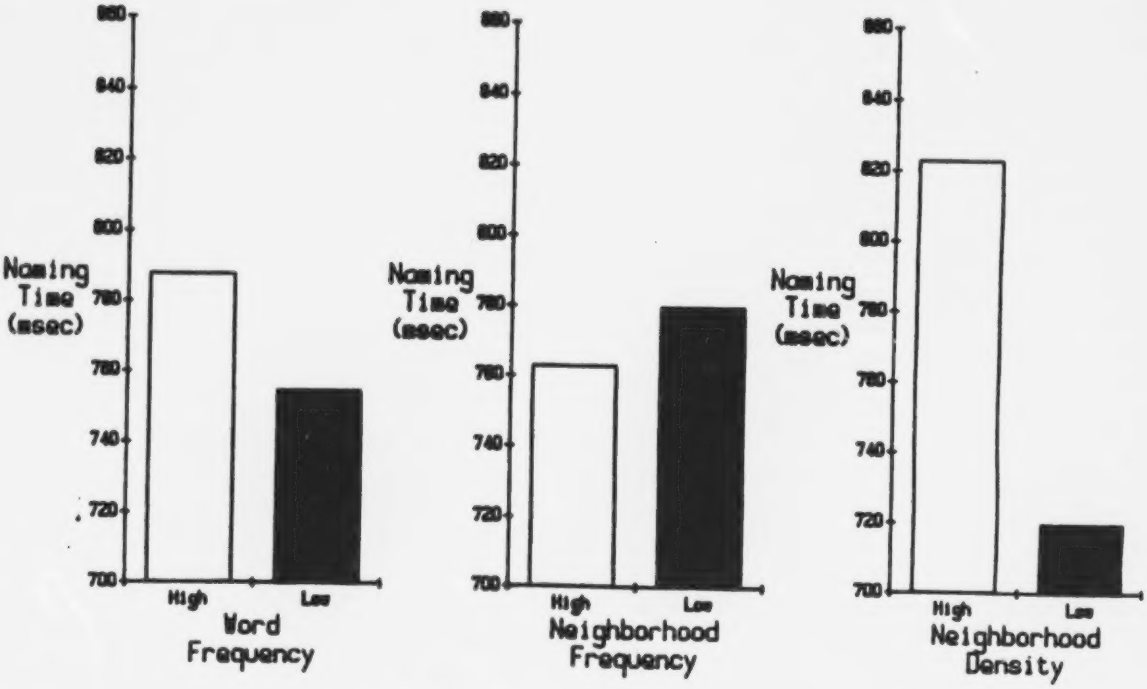


Figure 4. Mean naming response latencies for word frequency, neighborhood frequency, and neighborhood density.

stimulus word frequency or neighborhood frequency were observed.

As hypothesized, significant effects of density were observed in the absence of frequency effects. These results demonstrate that the naming task requires multiple activation of lexical items in memory, but that word frequency information that may bias decision processes operating on these units is bypassed. Because the naming response requires a precise analysis of the acoustic-phonetic properties of the stimulus word in order to build an articulatory plan for executing a response, biases not based on the acoustic-phonetics themselves (e.g., frequency biases) may hinder response generation. Given the response required by the naming task, therefore, subjects may optimize performance by focusing on discriminating among the acoustic-phonetic patterns and ignoring higher-level lexical information. Thus, frequency effects would not be expected to affect naming times. However, because an acoustic-phonetic pattern must be isolated in order to make the naming response, neighborhood density would be expected to influence the time need to generate the response. Indeed, precisely this pattern of results was obtained in the naming study.

The results of this study therefore demonstrate that neighborhood density effects are clearly separate from frequency effects. Thus, the present study demonstrates that stimulus similarity and decision biases based on frequency are separable factors that have differential effects on the levels of processing in the word recognition system.

Evidence from Primed Auditory Word Identification

The picture that begins to emerge from the three previous studies is one involving two processes in spoken word recognition: Activation of multiple candidates in memory and frequency-biased decision processes. Furthermore, these studies demonstrate that activation of multiple lexical items engenders costs to the word recognition system in terms of speed and accuracy of processing.

Neither of these proposals is consistent with at least one popular theory of spoken word recognition, namely cohort theory. The most recent version of cohort theory predicts neither processing costs associated with activation of multiple items nor frequency-biased decision processes. On this latter issue, cohort theory assumes that frequency adjusts the activation levels of items in memory and is not the product of post-activation decision processes.

In order to further investigate the issues of multiple activation and frequency in spoken word recognition, Goldinger, Luce, and Pisoni (1988) recently conducted a priming experiment aimed at evaluating two crucial aspects of the effects of similarity neighborhood structure on auditory word recognition. The first concerned predictions based on the neighborhood probability rule of inhibition from priming in spoken word recognition. The second concerned the issue of the separation of activation and frequency bias.

The neighborhood probability rule discussed earlier makes an interesting and somewhat counterintuitive prediction regarding priming by phonetically related items. Typically, priming effects in both visual and auditory word recognition experiments have demonstrated facilitation from the prime to the target item. However, the neighborhood probability rule predicts that in certain instances, inhibition should arise when a phonetically related prime is presented prior to a target item.

Recall that the rule states that identification performance may be characterized by the probability of choosing the stimulus word from among its neighbors. As the number and frequency of the neighbors increase, identification performance is predicted to decrease. Consider now the case in which a neighbor of the stimulus word is present prior to the stimulus word itself and subjects are required to identify the stimulus word. Assuming that some residual activation from the phonetically-related prime (the neighbor) is present upon presentation of the stimulus word, this residual activation from the prime should increase the overall activity of words in the neighborhood, thus reducing accuracy of identification of the stimulus word itself, relative to an appropriate baseline condition (i.e., phonetically-unrelated prime-stimulus word pair). That is, priming with a neighbor should actually result in reduced identification of the stimulus word, given that residual activation from the neighbor prime will result in increased competition with the stimulus word for identification.

In addition, it was predicted that the degree of inhibition would vary as a function of the frequency of the prime. In particular, it was predicted that low frequency primes should produce more inhibition than high frequency primes. The rationale for this prediction is as follows: All things being equal, low frequency words should be identified less quickly and less accurately than high frequency words. Recall that this prediction is not based on the assumption that high frequency words have higher resting activation levels, lower recognition thresholds, or steeper activation functions than low frequency words. Instead, the word frequency advantage is assumed to arise because biased decisions regarding the stimulus input can be made more quickly and accurately for high frequency words. Because frequency affects the decision process and not activation levels, at the moment a decision has been made regarding the identity of the stimulus input, activation levels of high frequency words will be lower than the activation levels of low frequency words. Simply stated, frequency-biased decisions about high frequency words are accomplished quickly, requiring less time for activation levels to rise for high frequency words than for low frequency words. Therefore, NAM predicts that low frequency primes will leave higher residual activation levels in the neighborhood when the target is presented, thus producing more inhibition than high frequency primes. Again, the higher residual activation for low frequency primes is simply a consequence of delayed frequency-biased decisions.

In order to test these two hypotheses, Goldinger et al. (1988) generated consonant-vowel-consonant target items varying in stimulus word frequency and neighborhood density and frequency. These target items were then paired with phonetically-related primes that were either high or low in frequency. The phonetically-related primes constituted the nearest neighbors of the target words that had no identical overlapping phonemes. Computation of primes from these neighborhoods was identical to that used in the perceptual identification study discussed above. The restriction that none of the primes have any overlapping phonemes was imposed to guard against guessing strategies based on subjects' expectancies of overlap. In addition to phonetically related primes, unrelated primes for each of the target stimuli were also generated. The unrelated primes had a probability of zero of being a neighbor of the target item.

On a given trial, a phonetically related or unrelated prime was presented in the clear immediately preceding presentation of a target item mixed with white noise at a +5 dB signal-to-noise ratio. The subjects were required to identify the target word on each trial.

The results of the primed auditory word identification study are shown in Figure 5. Light bars indicate conditions for neutral primes, dark bars show performance for related primes. Mean percentages of correct target identification for high frequency targets are shown on the left; mean percentages for low frequency targets are shown on the right. Mean percentages correct for prime-target pairs occurring in sparse neighborhoods are shown in the upper panel; percentages for dense neighborhoods are shown in the lower panel.

Insert Figure 5 about here

Significant effects of target word frequency and neighborhood density were obtained, replicating the effects of the previously discussed perceptual identification study. In addition, significant effects of inhibition priming were obtained in three conditions, indicated by asterisks in Figure 5. Note that significant effects of inhibition were obtained only for low frequency primes, as predicted, with one exception. No inhibition was observed for low frequency words occurring in high density neighborhoods. The failure to observe a significant effect may be due to the fact that in this condition, the target item is already receiving a great deal of competition from other words in its neighborhood. Thus, neighborhood competition may be so strong in this condition that any additional competition provided by the related prime preceding the target is undetectable. Nevertheless, as predicted by the neighborhood probability rule, related low frequency primes in all other conditions resulted in significant inhibition. These results demonstrate that phonetically related primes selected from the same similarity neighborhoods increases neighborhood competition via their residual activation, thus lowering the probability of correct selection of target items from among their neighbors. In addition, only low frequency primes produced significant levels of inhibition. This result is problematic for models that assume that frequency is coded in activation levels or thresholds. However, these results are easily accounted for by a model that separates

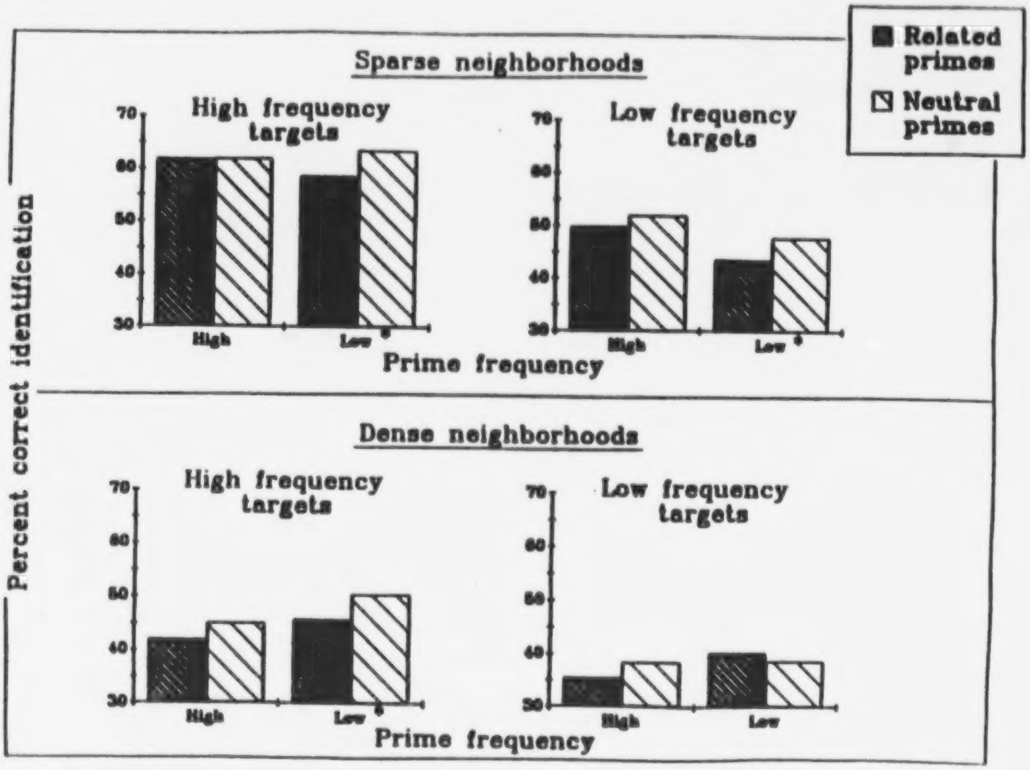


Figure 5. Results from primed auditory word recognition. Ligh bars indicate conditions for neutral primes, dark bars show performance for related primes.

activation and frequency-biased decision processes.

The Neighborhood Activation Model

On the basis of these and other results (see Luce, 1986), we have proposed a model of spoken word recognition that attempts to characterize the processes of neighborhood activation and selection. This model, called the neighborhood activation model, is primarily a processing instantiation of the frequency-weighted neighborhood probability rule described above. Basically, the model assumes that a set of similar acoustic-phonetic patterns is activated in memory on the basis of stimulus input. The activation levels of these patterns are assumed to be direct function of their phonetic similarity to the stimulus input. Over the course of processing, stimulus input serves to resolve or refine a pattern. That is, as processing proceeds, the pattern corresponding to the stimulus input receives successively higher levels of activation, while the activation levels of similar patterns are attenuated.

Words emerge in the neighborhood activation model when a system of word decision units tuned to the acoustic-phonetic patterns are activated. The activation of the decision units is assumed to be direct, in the sense of logogen theory (Morton, 1979) and cohort theory (Marslen-Wilson & Welsh, 1978). In addition, as in cohort theory, the system of word units is assumed to be based only on the activation of the acoustic-phonetic patterns. That is, word recognition is assumed to be, at least initially, completely data driven based on information contained in the speech waveform. Once the word decision units are activated, they monitor a number of sources of information. The first source of information is the activation of the acoustic-phonetic patterns, which have previously served to activate the decision units themselves. The word decision units also monitor the overall level of activity in the decision system itself, much like processing units monitor the net activity level of the system in the TRACE model of speech perception (Elman & McClelland, 1986; McClelland & Elman, 1986). Finally, decision units are tuned to higher-level lexical information, which includes word frequency. This information serves to bias the decisions of the units by weighting the activity levels of the words to which they respond. The values that serve as the output of the decision units are assumed to be computed via a rule similar to the neighborhood probability rule discussed above.

Word recognition in the neighborhood activation model may be accomplished in a number of ways, depending on the requirements of the task. In situations in which the stimulus input is degraded, word recognition is accomplished by evaluating the values computed by the decision units and selecting a response based on these values. When speeded responses are required, it is assumed that the subject sets a criterion for responding that, once exceeded by the output of a decision unit, results in the recognition of a word. Word recognition is defined explicitly as the choice of a particular pattern by the system of decision units. Lexical access is assumed to occur once a decision unit makes all of the information it was monitoring available to working memory. Thus, the decision units act as gates on the acoustic-phonetic and lexical information available to the processing system. If insufficient evidence for a word is provided by the decision system, the activation levels of the acoustic-phonetic patterns themselves may be consulted, resulting in the recognition of a nonword pattern.

The neighborhood activation model places much of the burden of spoken word recognition on the discrimination among similar acoustic- phonetic patterns corresponding to words and the decisions necessary for choosing among these patterns. In addition, the model accounts for word frequency effects by allowing frequency information to bias the decisions of the word decision units. However, because each word decision unit computes values based on its acoustic-phonetic pattern as well as the overall level of activity in the decision system, decisions are assumed to be context sensitive. Thus, frequency is assumed to be a relative factor, not an inherent property or attribute of a word. That is, if many decision units are receiving strong frequency biases, a decision unit for a given high frequency word may compute relatively low values. Likewise, a decision unit for a low frequency word may quickly begin to output high values if there is little other activity in the system. Thus, effects of frequency are not assumed to be absolute, but instead are highly dependent on the activity level of the decision system as a whole.

Other Models of Word Recognition

The neighborhood activation model bears a strong resemblance to other models of spoken word recognition, and many of the concepts incorporated in the model have precedents in previous accounts of auditory word recognition. However, as will be argued below, the model makes certain predictions that are inconsistent with current models of word recognition, in particular with regard to the roles of frequency and similarity. We now turn to a discussion of two of the more influential models of word recognition in order to highlight the fundamental differences and similarities between the neighborhood activation model and these models.

Logogen Theory

Morton (1969, 1979) has proposed a model of word recognition based on a system of "logogens" that monitor bottom-up sensory information and top-down contextual and lexical information. Information from either of these sources serves to drive the logogens toward threshold. Once a threshold is reached, the information to which the logogen corresponds is made available to the processing system and a word is said to be recognized and accessed. Morton accounts for word frequency effects in the logogen model by assuming that high frequency words require less evidence than low frequency words to cross threshold. Morton thus refers to logogen theory as an evidence-bias model.

The resemblance between Morton's system of logogens and the system of word decision units in the neighborhood activation model is quite strong. Both logogens and word decision units monitor top-down and bottom-up information. In addition, both logogens and word decision units are assumed to prohibit information from becoming available to the general processing system until a decision regarding the identity of the word has been made. However, word decision units differ from logogens in a number of important ways.

Perhaps the most crucial difference between logogens and the word decision units hinges on the problem of accounting for neighborhood structural effects. Logogens are assumed to be independent processing units with no interconnections among lexical items in memory. The lack of crosstalk among logogens makes it difficult to account for the findings that words in highly dense or confusable neighborhoods take longer to respond than words in less dense or less confusable neighborhoods. Because logogens are independent processing units, stimulus input should push a given logogen over threshold at the same point in time, regardless of whether the stimulus input activates many or few logogens. Granted, accuracy differences between dense and sparse neighborhoods may arise because there is a higher probability that logogens corresponding to similar words may surpass threshold prior to the logogen corresponding to the stimulus input. It is not so clear, however, how logogen theory would account for the effects of neighborhood density on reaction times. When presented with clearly specified acoustic-phonetic information, as in auditory word naming, the logogen corresponding to the stimulus input should always cross threshold at the same point in time, regardless of the activity levels of other logogens, assuming that word frequency is held constant.

The most difficult problem that the present set of results poses for logogen theory concerns the robust findings that frequency effects are dependent on the neighborhood structure of the stimulus word. In the perceptual identification study, it was shown that certain classes of high and low frequency words are responded to at equal levels of accuracy if the neighborhood structures of the words were equated. Because logogens corresponding to high and low frequency have different thresholds, low frequency words should always require more evidence than high frequency words in order to cross threshold. Because a single logogen has no knowledge of the activation levels of other logogens, it is difficult to explain within logogen theory how the frequencies of items in a neighborhood could influence recognition of the stimulus word.

In addition, logogen theory has no mechanism for explaining the results of the word naming study. Recall that in the naming study we argued that word units must have been accessed by subjects in order to produce the effect of neighborhood density. However, no effects of word frequency or neighborhood frequency were observed. It is perhaps possible that the thresholds for logogens corresponding to high and low frequency words were temporarily equated due to some unspecified property of the naming task. However, not only is this solution extremely inelegant and unparsimonious, it seriously calls into question logogen theory's fundamental claim that thresholds are intrinsic to the logogens themselves and arise over time as a function of degree of exposure to words.

A final problem for logogen theory concerns the nonword data from the auditory lexical decision experiment. Coltheart, et al., (1976) have proposed that nonword decisions in the logogen model can be made in a similar manner to nonword decisions in the neighborhood activation model. Specifically, a nonword decision is executed when no logogen fires. However, because the activation levels within the logogens are not available for inspection (i.e., logogens are either above or below threshold), it is difficult to account for the finding that the number and nature of words activated by the nonword stimulus influences reaction time. As logogen theory stands, there is no means for evaluating the overall level of activity in the logogen system, and consequently there is no mechanism for making faster decisions to nonwords with fewer neighbors or lower frequency neighbors. The nonword data from the auditory lexical decision experiment are therefore problematic for a system of independent processing units that respond only upon surpassing an intrinsic thresholds.

The neighborhood activation model, on the other hand, provides a coherent description of the present set of results by assuming that the decision units are interconnected and that frequency effects arise from biases stemming from higher-level sources of information. Modifications of logogen theory may be possible to account for the present results, but it is very likely that the resulting model would bear a strong resemblance to the neighborhood activation model. Nonetheless, there are important similarities between the neighborhood activation model and logogen theory, owing to the fact that the present model incorporates many ideas from logogen theory. In particular, the neighborhood activation model assumes a system of word decision units that serve as the interface between the acoustic- phonetic input and higher-level information, as proposed by logogen theory. However, due to the interconnectedness of the system of word decision units, the neighborhood activation model is able to account for the effects of neighborhood structure, whereas logogen theory apparently is not.

Cohort Theory

Perhaps the most influential of current models of spoken word recognition is cohort theory, proposed by Marslen-Wilson (Marslen-Wilson & Welsh, 1978; Marslen-Wilson & Tyler, 1980; Marslen-Wilson, 1984, 1987). According to this theory, a "cohort" of words is activated in memory on the basis of the initial acoustic-phonetic input of the stimulus word. Words in the cohort are then eliminated by two sources of information: continued acoustic-phonetic input and top-down contextual information. That is, words in the cohort are ruled out or deactivated by continued processing of the stimulus information as well as by inconsistent contextual information. A given word is recognized when it is the only word remaining in the cohort.

Cohort theory has provided a number of valuable insights into the temporal processing of spoken words. In previous versions of the theory, however, no attempt was made to account for word frequency effects. In a recent version of the theory, though, Marslen-Wilson (1987) has incorporated a mechanism for accounting for word frequency effects by assuming that words in a cohort have differing levels of activation depending on their frequencies of occurrence. Words with higher levels of activation take longer to eliminate from the cohort than words with lower levels of activation, thus affording at least an initial advantage to high frequency words. Because the latter version of cohort theory represents a significant improvement over the initial formulation of the theory, only this version will be considered in the present discussion.

Cohort theory and the neighborhood activation model are similar in the respect that both models assume bottom-up priority in the activation of items in memory. Furthermore, both models assume that a set of items is activated and processed in parallel. In addition, both models state that items receive reduced levels of activity as disconfirming acoustic-phonetic information is presented. Unlike cohort theory, however, the neighborhood activation model at this stage of formulation has little to say about the time course of effects in the word recognition system, primarily due to the fact that the model was developed on the basis of data from very short words. Indeed, some of the aspects of cohort theory may have to be incorporated into the neighborhood activation model in order to account for the recognition of longer words. Nonetheless, cohort theory and the present model do make fundamentally different predictions, at least for short stimuli.

Marslen-Wilson (1987) argues that because cohort theory is realized as a parallel system, no effects of set size should be observed on word recognition. Words in a cohort are assumed to be activated at no cost. The neighborhood activation model is also realized as a system of parallel processing units, but the fundamental claim of the neighborhood activation model is that the nature and number of items activated in memory does influence the accuracy as well as the speed of recognition. This prediction stems from the claim that the word decision units are sensitive to the overall level of activity in the decision system and are therefore influenced by the number and nature of competing items. Evidence to support this claim was provided by each of the experiments previously reported.

Marslen-Wilson (1987) argues that set size has no effect on recognition performance on the basis of a set of experiments examining lexical decisions for nonwords. He claims that if nonwords are matched at the point at which they diverge from words, no effect of set size should be observed on reaction times. This is in contradiction to the findings in the lexical decision experiment reported earlier in which large effects of neighborhood density (i.e., set size) and neighborhood frequency were observed for nonwords. Recall that because of the manner in which these nonwords were constructed, each of the nonwords diverged from words at the third phoneme. Thus, set size effects were demonstrated even when divergence points were equated. Given that Marslen-Wilson's claim of no effects of set size are based on null results, the positive findings reported for the nonwords seriously calls this claim into question.

Indeed, each of the experiments reported previously fails to support the notion that the number of items activated in memory has no influence on recognition performance. Although Marslen-Wilson may object to the results from the perceptual identification study, claiming that the use of "noisy " stimuli may induce post-perceptual processes, the results from the lexical decision study taken together with the auditory naming study clearly contradict a fundamental claim of cohort theory. Indeed, it is not even clear that the postulation of some vague "post-perceptual" processes indicts the results from the perceptual identification study, which showed significant effects of neighborhood structure on identification performance. In short, the results of the present set of studies considered together refute a fundamental claim of cohort theory that activation of multiple lexical items in memory results in no processing costs to the system.

The results of the naming study also provide counter evidence to cohort theory's treatment of word frequency. All words used in the naming study had approximately equal isolation points by virtue of their short length (see Luce, 1986a). However, despite equivalent isolation points, high frequency words were named no faster than low frequency words, in contradiction to the predictions made by the most recent version of cohort theory (Marslen-Wilson, 1987). In addition, because there was a strong effect of density, it cannot be assumed that lexical items were bypassed in the generation of the naming response. Thus, the current version of cohort theory also fails to account for the results obtained in the present investigation that frequency effects may be circumvented by task requirements.

Finally, the results of the priming study provide further evidence that high and low frequency words may not differ in absolute activation levels and that effects of frequency are at a stage following activation. This conclusion is supported by the finding that only low frequency primes resulted in significant levels of inhibition, a result that cannot easily be accounted for by models that assume that frequency is coded in activation levels or thresholds. As previously argued, an adequate model of spoken word recognition cannot assume differing inherent activation levels or thresholds for the units monitoring high and low frequency words. Instead, the effects of frequency are best described as biases on the decision units responsible for choosing among activated lexical items. By treating the effects of frequency as biases on the decision process, one can account for results demonstrating the lability of the frequency effect depending on task requirements (e.g., Pollack, et al., 1959) and higher-level sources of information (Grosjean & Itzler, 1984). Thus, the instantiation of frequency in the latest version of cohort theory is difficult to countenance. The neighborhood activation model, however, provides a more principled explanation of the effects of word frequency on both the stimulus word and its neighbors.

Conclusion

The picture that begins to emerge from the results reported in the studies discussed above is one of a perceptual and cognitive system optimized for the recognition of words under a variety of circumstances. This optimization is achieved by a simultaneous activation of alternatives based on the stimulus input, and by a sophisticated system that attempts to maximize decisions among these alternatives. The fact that the word recognition system is capable of considering numerous alternatives in parallel helps to assure the best performance of the system in the face of stimulus input that is often impoverished, degraded, or poorly specified. However, as the present set of experiments have shown, this optimization is not without its processing costs. Both the number and nature of words activated by the stimulus input affect not only the accuracy of word recognition, but also the time required to decide among the activated candidates. Nevertheless, such processing costs subserve the ultimate goal of the speech perception, namely to maximize the speed and accuracy with which words are recognized in real-time. In short, the study of the structural organization of the neighborhoods of words in the mental lexicon has provided deeper insights into one important aspect of the fundamentally and uniquely human capacity to communicate with spoken language.

References

Balota, D. A., & Chumbley, J. I. (1984). Are lexical decisions a good measure of lexical access? The role of word frequency in the neglected decision stage. *Journal of Experimental Psychology: Human Perception and Performance*, **10**, 340-357.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production frequency? *Journal of Verbal Learning and Verbal Behavior*, **24**, 89- 106.

Coltheart, M., Develaar, E., Jonasson, J. T., & Besner, D. (1976). Access to the internal lexicon. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, NJ: Erlbaum.

Elman, J. L., & McClelland, J. L. (1986). Exploiting lawful variability in the speech waveform. In J. S. Perkell & D. H. Klatt (Eds.), *Invariance and variability in speech processing*, (pp. 360- 385). Hillsdale, NJ: Erlbaum.

Goldinger, S. D., Luce, P. A., & Pisoni, D. B. (1988). Priming lexical neighbors of spoken words: Effects of competition and inhibition. Submitted for publication.

Grosjean, F., & Itzler, J. (1984). Can semantic constraint reduce the role of word frequency during spoken-word recognition. *Perception and Psychophysics*, **22**, 180-182.

Howes, D. H. (1957). On the relation between the intelligibility and frequency of occurrence of English words. *Journal of the Acoustical Society of America*, **29**, 296-305.

Landauer, T. K., & Streeter, L. A. (1973). Structural differences between common and rare words: Failure of equivalence assumptions for theories of word recognition. *Journal of Verbal Learning and Verbal Behavior*, **12**, 119-131.

Luce, P. A. (1986a). A computational analysis of uniqueness points in auditory word recognition. *Perception and Psychophysics*, **39**, 155-158.

Luce, P. A. (1986b). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6*. Bloomington, IN: Indiana University.

Luce, P. A. (1988a). Similarity neighborhoods of words in the mental lexicon: A computational analysis. Manuscript in preparation. Bloomington, IN: Indiana University.

Luce, P. A. (1988b). Stimulus context and similarity neighborhood structure. Manuscript in preparation.

Luce, R. D. (1959). *Individual choice behavior*. New York, NY: Wiley.

Marslen-Wilson, W. D. (1984). Function and process in spoken word recognition: A tutorial review. In H. Bouma and D. G. Bouwhuis (Eds.), *Attention and performance X: control of language processes*. Hillsdale, NJ: Erlbaum.

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word recognition. In U. H. Frauenfelder and L. K. Tyler (Eds.), *Spoken word recognition*. Cambridge, MA: MIT Press.

Marslen-Wilson, W. D., & Tyler, L. K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, **10**, 29-63.

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, **76**, 165-178.

Morton, J. (1979). Word recognition. In J. Morton and J. D. Marshall (Eds.), *Psycholinguistics 2: Structures and processes*, (pp. 107- 156). Cambridge, MA: MIT Press.

Newbigging, P. L. (1961). The perceptual redintegration of frequent and infrequent words. *Canadian Journal of Psychology*, **15**, 123-132.

Nusbaum, H. C., Pisoni, D. B., & Davis, C. K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10*. Bloomington, IN: Indiana University.

Paap, K. R., McDonald, J. E., Schvaneveldt, R. W., & Noel, R. W. (1987). Frequency and pronounceability in visually presented naming and lexical-decision tasks. In M. Coltheart (Ed.), *Attention and performance XII: The psychology of reading*. Hillsdale, NJ: Erlbaum.

Pollack, I., Rubenstein, H., & Decker, L. (1959). Intelligibility of known and unknown message sets. *Journal of Acoustical Society of America*, **31**, 273-259.

Savin, H. B. (1963). Word-frequency effect and errors in the perception of speech. *Journal of the Acoustical Society of America*, **35**, 200-206.

Soloman, R. L., & Postman, L. (1952). Frequency of usage as a determinant of recognition thresholds for words. *Journal of Experimental Psychology*, **43**, 195-201.

Triesman, M. (1978a). A theory of the identification of complex stimuli with an application to word recognition. *Psychological Review*, **85**, 525-570.

Triesman, M. (1978b). Space or lexicon? The word frequency effect and the error response frequency effect. *Journal of Verbal Learning and Verbal Behavior*, **17**, 37-59.

Webster's Seventh Collegiate Dictionary. (1967). Los Angeles: Library Reproduction Service.

Manner of Articulation and Feature Geometry: A Phonological Perspective[1]

Stuart Davis

Speech Research Laboratory
Department of Psychology
Indiana Uiversity
Bloomington, IN 47405

---

[1] This research was supported by NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN. Thanks go to Keith Johnson for discussion on this paper.

Abstract

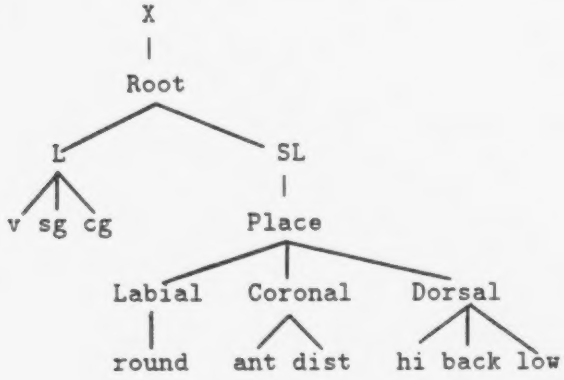
Recent work in phonological theory has argued for a view that the features that comprise a phoneme are hierarchically organized. The conception of phoneme feature organization that has emerged from both a phonetic and phonological perspective is that, on the one hand, place of articlulation features (e.g., [back], [high], [anterior], etc.) group together while, on the other hand, laryngeal features (e.g., [voice], [spread glottis], [constricted glottis]) group together. Moreover, among the place features further organization has been posited depending on which articulator controls the feature. For example, the tongue body controls (or articulates) the features [high], [back], and [low]. These are referred to as dorsal features and they group together. The tongue blade controls the features [anterior] and [distributed]. These are referred to as coronal features and they group together. And the lips are used in articulating the feature [round] which is a labial feature. Thus the place of articulation features are further divided into dorsal, coronal, and labial features. While there is fairly widespread consensus among phonologists on the general division into laryngeal features and place features, there is very little consensus on the status of manner of articulation features (e.g., [nasal], [continuant], [lateral]) within Feature Geometry. This paper focuses on the issue of just how the manner feature [continuant] fits into the hierarchical conception from a phonological perspective. Based on a relatively small range of phenomena it is argued that the feature [continuant] is best represented as a supralaryngeal feature.

Manner of Articulation and Feature Geometry: A Phonological Perspective

Introduction

Much recent work in phonological theory has focused on the internal organization of phonological features. From the works of such researchers as Clements (1985,1987), Archangeli and Pulleyblank (1986), Halle (1986), and Sagey (1986b) a view has emerged that phonological features are hierarchically organized with certain features grouping together. This emerging view of the organization of phonological features (i.e., of Feature Geometry) can be represented by a hierarchical tree structure.[1] This is shown in (1), below.[2]

(1)



The organization in (1) reflects the view that phonological features (indicated with small case letters) can be divided into supralaryngeal and laryngal features. These features are represented in (1) as being under the Supralaryngeal and Laryngeal node, respectively. The supralaryngeal features would include place of articulation features. These place of articulation features, as indicated in (1), are divided into tongue body features (i.e., dorsal features), tongue blade features (i.e., coronal features), and labial features. In the representation in (1) the Laryngeal node and the Supralaryngeal node are dominated by the Root node.[3] The

---

[1] The following abbreviations are used: L=laryngeal, SL=supralaryngeal, v=voice, sg=spread glottis, cg=constricted glottis, ant=anterior, cor=coronal and X is a C-slot or V-slot on the CV-tier or timing tier.

[2] We ignore here the issue of the location of manner of articulation features. This issue will be addressed in the following section.

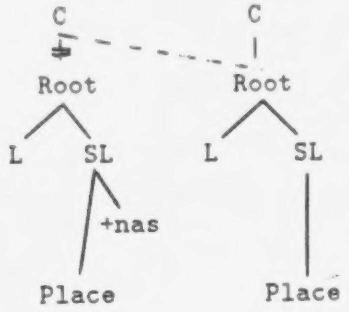
[3] The nonterminal nodes in (1), Root, Laryngeal, Supralaryngeal, and Place are referred to by Clements (1985) as class nodes. In the numbered examples throughout this paper class nodes will begin with a capital letter so as to distinguish them from phonological features which, in the hierarchical tree structure like in (1), are terminal nodes immediately dominated by class nodes.

Root node links directly to a C-slot or V-slot on the CV-tier.[4]

Both phonetic and phonologocial evidence have been adduced for the feature organization in (1). The phonetic evidence, as discussed in Clements (1985) (and paraphrased here), comes from the observation that speech involves the coordination of simultaneous and partially overlapping laryngeal and supralaryngeal gestures. That is, in speech one can keep the laryngeal configuration constant while varying the geometry of the vocal tract, or one can keep the supralaryngeal gesture constant while varying the state of the larynx. The division of features into laryngeal and supralaryngeal, as in (1) captures this observation. Moreover, regarding place of articulation features, the labial features, the coronal features, and the dorsal features are each independent of one another, so that it is quite possible to articulate a labiovelar (which involves both labial and dorsal features) or a labiocoronal (which involves both labial and coronal features), but it is impossible to articulate a velar-uvular since both velars and uvulars are made with the tongue body.

The phonological evidence for the hierarchical structure in (1) is discussed at great length in Clements (1985) and Sagey (1986b). Basically, they show that rules of assimilation and deletion often involve just class nodes. For example, rules of total assimilation involve Root node spreading. This is shown in (2) for a rule that totally assimilates a nasal consonant to the following consonant (nas=nasal). [5]

(2)



Moreover, rules in which there is deletion or assimilation of just laryngeal features, supralaryngeal features, or place features would involve the delinking or spreading of the Laryngeal, Supralaryngeal, or Place nodes, respectively. Consider, for example, the process in the Native American language Klamath whereby a lateral ejective becomes a glottal stop

---

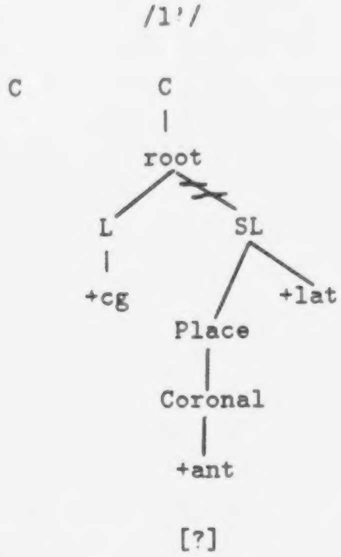
[4] We assume in this paper the theory of CV phonology as put forth in Clements and Keyser (1983). Nothing in this paper crucially hinges on this particular theory as opposed to the X-tier theory of Levin (1985).

[5] We assume here that assimilation phenomena are more insightfully analyzed as being accomplished through autosegmental spreading rather than through feature changing rules. See Hayes (1986) for specific argumentation for the superiority of the autosegmental account of assimilation.

184

[?] after certain consonants. Under a view of Feature Geometry such as in (1), this phenomenon can be explained as simply the loss (or delinking) of all supralaryngeal features of the lateral ejective while its laryngeal features remain intact (lat=lateral). This is shown in (3).

(3)

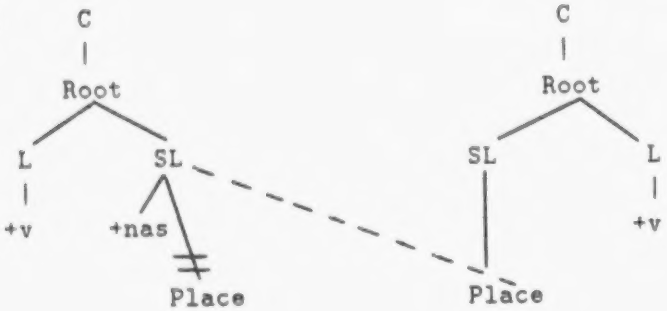


In addition to the evidence from total assimilation and deletion phenomena, partial assimilation processes also provide evidence for the hierarchical feature representation in (1). This is discussed at length by Clements (1985) and Sagey (1986b). They note that in language after language assimilation rules involve certain combinations of features but not others. For example, the features [anterior], [coronal], [high], and [back] often are mentioned together in assimilation rules, particularly in rules whereby nasal consonants assimilate to the place of articulation of a following consonant. Such a rule would be written as in (4) using the rule-writing formalism of Generative Phonology.

```
(4)  C                              C
    +nas ----> a ant   / ___   a ant
               b cor           b cor
               c high          c high
               d back          d back
```

Under the account in (4) there is no particular reason why it is the features [anterior], [coronal], [high], and [back] that assimilate together and not, for example, [back] and [voice]. However, on an analysis of assimilation that incorporates a hierarchical representation of phonological features as in (1) (as well as viewing assimilation as being accomplished by autosegmental spreading), the reason why it is quite common for the combination of features [anterior], [coronal], [high], and [back] to assimilate together is that such cases are just instances of place of articualation assimilation. A rule like that in (4), on a feature geometry account, simply involves the spreading of the Place node. This is seen below.

(5)



An account of assimilation that incorporates Feature Geometry would predict that certain types of assimilation would be nonoccurring (or, at least, extremely uncommon) such as a rule that assimilates the features [back] and [voice] together, as in (6). Such a rule might, for example, convert /g/ to [k^y] in the environment before a /p/.

(6)

$$C \dashrightarrow \begin{array}{l} \text{a back} \\ \text{b voice} \end{array} \; / \; \_\_\_ \; \begin{array}{l} \text{C} \\ \text{a back} \\ \text{b voice}. \end{array}$$

Rules like this would be predicted to be nonoccurring (or, at least, extremely uncommon) because the features [voice] and [back] do not group together. The feature [voice] is a laryngeal feature dominated by the Laryngeal node (in a representation like that in (1)) and the feature [back] is a place of articulation feature dominated by the Dorsal node. Under a standard generative phonology account of assimilation it would be predicted that a rule like that in (6) would be more common than a rule like that in (4) since fewer features are involved in (6). The fact that assimilations like that in (6) are either nonoccurring (or, extremely uncommon) while assimilations like that in (4) are quite common thus constitutes evidence for the hierarchical organization of phonological features along the lines in (1).[6]

In the current work on Feature Geometry there is widespread agreement that the Root node dominates both the Supralaryngeal and Laryngeal nodes, and the Supralaryngeal node dominates the Place node.[7] In turn, the Laryngeal node dominates features related to the larynx while the Place node dominates at least the Labial, Coronal, and Dorsal Nodes. What is very uncertain in the current work on Feature Geometry is how the manner of articulation features (e.g., [continuant], [consonantal], [strident], etc.) fit in. So far in the emerging literature on Feature Geometry there has been precious little argumentation either phonetically or phonologically that has direct bearing manner of articulation features. In the remainder of this paper we focus our attention on manner features and specifically on the feature [continuant]. In the following section, we critically review the various proposals concerning the location of the feature [continuant], and, afterwards, we argue based on phonological phenomena found in several different languages that the feature [continuant] is best represented as a supralaryngeal feature.

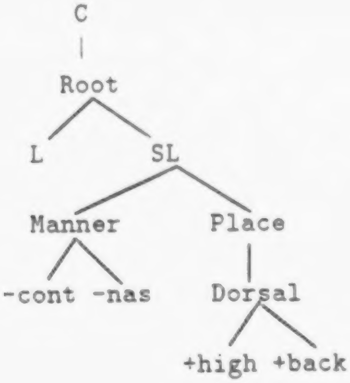
---

[6] It should be pointed out that assimilations may also involve single features. In a theory that incorporates a hierarchical representation of features (like that in (1)), such assimilations would simply involve autosegmental spreading of that individual feature and not necessarily of a class node like in (5).

Proposals Concerning [Continuant]

In the current work on Feature Geometry there are at least two views concerning the location of the feature [continuant]. One view, held by Clements (1985), Archangeli and Pulleyblank (1986) and Clements (1987) considers the feature [continuant] to be a supralaryngeal feature, and, thus, in a formal representation like that in (1), it would be located under the Supralaryngeal node. These researchers, though, have somewhat different views on exactly where under the Supralaryngeal node the feature [continuant] should be located. Clements (1985) groups the feature [continuant] together with other manner of articulation features as a single block off of the Supralaryngeal node, as shown for the representation of /k/ in (7). Note that specific features will be indicated only when directly relevant ([cont]=continuant]).

(7)



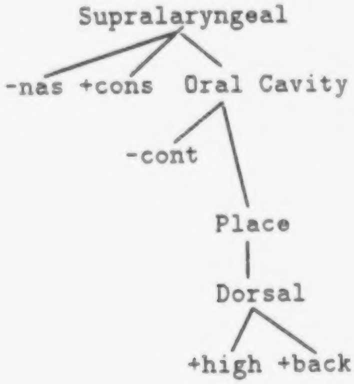
More recently, though, Clements (1987) has argued that the feature [continuant] is not grouped together with the other manner features but rather is grouped by itself off of the

---

[7] For specific argumentation for the Laryngeal node and the Supralaryngeal node, see Clements (1985) and Sagey (1986b).

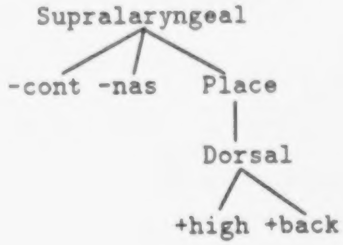
Oral Cavity node. Clements's proposed Oral Cavity node dominates the feature [continuant] and the Place node. The Oral Cavity node in turn is linked to the Supralaryngeal node. This view of the Supralaryngeal node is illustrated in (8) where the supralayngeal features of /k/ are shown. The other manner features, in addition to [continuant], would each be individually linked directly to the Supralaryngeal node.

(8)



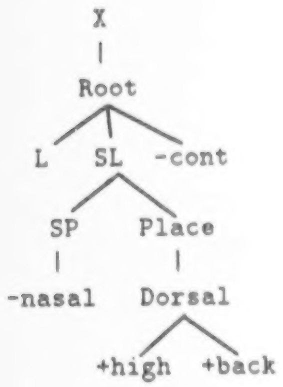
A different understanding of the exact location of the feature [continuant] under the Supralaryngeal node is assumed by Archangeli and Pulleyblank (1986). They take up a suggestion in Clements (1985), although not incorporated by him, that each manner feature (including [continuant]) is individually linked directly to the Supralaryngeal node. Their representation of the structure under the Supralaryngeal node for the phoneme /k/ would be as in (9).

(9)



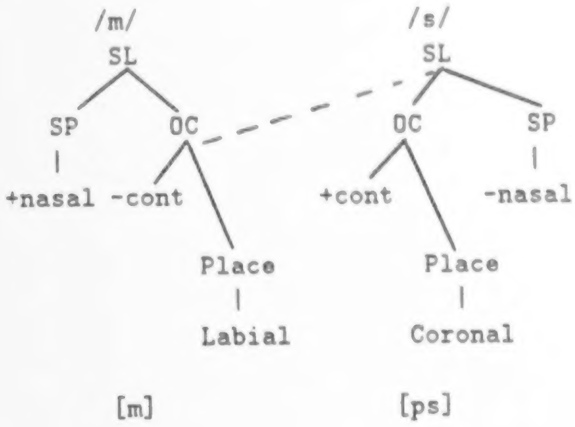
The second view of the location of the feature [continuant] is put forth in Sagey (1986b) where she argues that degree of closure features like [continuant] and [consonantal] are represented as being directly linked to the Root node; they are not really supralaryngeal features at all. Her representation of the phoneme /k/ would be as in (10). Note that Sagey (1986b) posits a Soft Palate node (SP) which dominates only the feature [nasal]. This difference will not concern us here.

(10)



Both the view of Clements (1985) that [continuant] groups together with other manner features as in (7) and the view adopted by Archangeli and Pulleyblank (1986) that [continuant] is linked directly to the Supralaryngeal node as in (9) have not been argued for. Neither of these authors present any evidence to support their view of the location of [continuant]. They do not seem to be aware of either spreading or delinking rules that would support their positions. Clements (1987), on the other hand, argues for the Oral Cavity node (i.e., the grouping of place features and [continuant] under a single node as in (8)) based mainly on an analysis of English intrusive stop formation. Clements argues that the intrusive [p] in American English words such as *hamster*, it warmth, *triumph*, and *dreamt*, the intrusive [t] in such words as *sense*, *ninth*, *censure*, *false*, *health*, and *Welsh*, as well as the intrusive [k] in such words as *youngster*, *length*, and *anxious* can be accounted for by means of autosegmental spreading of the Oral Cavity node of the nasal or lateral to the Supralaryngeal node of the following voiceless obstruent. This is illustrated in (11) where the /ms/ cluster from the word *hamster* is realized as [mps]. (OC=oral cavity)

(11)



Clements's solution, in which phonetically there is a lag of the oral cavitiy features, differs from the traditional one in which the intrusive stop comes about due to an anticipation of the orality and voicelessness of the following obstruent (see, for example, Anderson (1976).[8] Clements, however, criticizes the traditional analysis in that it is not articulatorily accurate. Clements cites the study of Ali, Daniloff, and Hammarberg (1979) which shows that it is the delayed release of the oral occlusion that brings about the intrusive stop and not anticipatory velar raising. Thus Clements's analysis of English intrusive stops, illustrated in (11), in which the Oral Cavity node is posited appears articulatorily more adequate than the traditional one.

Clement's argument for the Oral Cavity node based on the rule of Intrusive Stop Formation appears solid. One problem, however, with the notion of an Oral Cavity node is that one would expect to find cases of oral cavity assimilation. That is, if the Oral Cavity node is indeed part of the phonological representation, as Clements argues, there should be cases of assimilation that involve the continuant and place features of one phoneme assimilating to the continuant and place features of a second phoneme. For example, one would expect rules like those in (12), which involve the assimilation of [continuant] and place features, to be common.

(12) a. f ---> t / __ n,l

b. s ---> k / __ g

c. p ---> s / __ z

d. g ---> v / __ f

But assimilations like those in (12) are probably nonoccurring, or at best extremely rare. It appears, then, that the only motivation for the Oral Cavity node comes from intrusive stops since there does not seem to be additional evidence outside of intrusive stops for the Oral Cavity node. It may well be that Intrusive Stop Formation involves two different spreading processes (i.e., [-cont] spreading and place node spreading) occurring independently of one another. If such is the case, the data from the formation of intrusive stops does not provide conclusive evidence on where the feature [continuant] is located; since, if the spreading of the feature [-continuant] is separate from the spreading of the Place node, then, it is possible for the feature [continuant] to be located under just about any node.

---

[8] In Clements's analysis of English intrusive stops there is no insertion of a C-slot since various studies (e.g., Fourakis & Port (1986)) have shown that intrusive stops are of shorter duration than nonintrusive ones.

Sagey (1986b) supports her view that the feature [continuant] is located off of the Root node (shown in (12)) by evidence from complex segments (such as labiovelars). She shows that [continuant] could not be located under the Place node as in Sagey (1986a), for positing such would wrongly entail that place assimilation also involves [continuant] assimilation. She then suggests that degree of closure features (i.e., [continuant] and [consonantal]) are immediately dominated by the Root node. Her argument, though, is not incompatible with the view that [continuant] is immediately dominated by the Supralaryngeal node. And, in fact, she does not argue against the possibility that [continuant] is located under the Supralaryngeal node.

Thus, we find that in the current work on Feature Geometry there is a general lack of solid evidence on where the feature [continuant] is located. Of the possibilities shown in (7)-(10) none of them can really be ruled out. Perhaps the possiblity in (7) can be ruled out since cases where several manner features assimilate together in a single assimilation rule are virtually unattested in the literature. (But see Cho (1988) for a possible case in Korean.) The possibility in (8) finds support from the formation of intrusive stops in English as Clements (1987) has shown, but it does not seem to find any support from other rule types. Perhaps future research will show conclusively whether the posited Oral Cavity node can find support outside of the intrusive stop evidence. Hence, for now, we are left with two possibilities on the location of the feature [continuant]. It is either linked directly to the Root node as in (10) or it is linked directly to the Supralaryngeal node as in (9). We will henceforth refer to the possibility in (10), where the feature [continuant] is directly linked to the Root node, as Theory R (mnemonic for Root node); and we will refer to the possibility in (9), in which [continuant] is directly linked to the Supralaryngeal Node as Theory SL (mnemonic for Supralaryngeal node).

In the following section of this paper I examine several cases of phonological processes that involve the feature [continuant]. The phenomena to be considered are: affricate assimilation in the Dravidian language of India Pengo (which is to be examined in great detail), glottalization in Kinyarwanda (a Bantu language of Africa) and in English, and [continuant] deletion in Basque. Based on this range of phenomena it will be argued that from a phonological perspective [continuant] is best considered as a supralaryngeal feature.

The Location of the Feature [continuant]

Pengo Affricate Assimilation

In this subsection, I show that the facts of affricate assimilation in Pengo require that the feature [continuant] be located off of the Supralaryngeal node. At first glance it seems quite possible to formulate a concise analysis of the Pengo data that has the feature [continuant] located immediately under the Root node. However, when a wider range of data are considered it becomes apparent that only an analysis in which [continuant] is located off of the Supralaryngeal node is able to handle the Pengo data adequately. Thus Pengo Affricate Assimilation provides strong evidence for Theory SL.

The representative data that are to be considered first are presented in (13)-(16). These data, as well as all the other Pengo data referred to in this paper, are taken from Burrow & Bhattacharya (1970).[9] (ptvs=past tense verb stem)[10]

(13)	a.	/kuc-teng/	[kucceng]	'to sit'
	b.	/krac-teng/	[kracceng]	'to excrete'
	c.	/gac-teng/	[gacceng]	'to bind'
	d.	/muc-teng/	[mucceng]	'to bury'
	e.	/pac-t/	[pacc-]	'scratch' (ptvs)
	f.	/jo:c-t/	[jo:cc-]	'carry on the head' (ptvs)
(14)	a.	/uj-deng/	[ujjeng]	'to suck'
	b.	/a:nj-deng/	[anjeng]	'to catch'
	c.	/ne:nj-deng/	[ne:njeng]	'to breathe'
	d.	/vanj-deng]	[vanjeng]	'to cook'
(15)	a.	/kic-deng/	[kijjeng]	'to pinch'
	b.	/ho:c-deng/	[ho:jjeng]	'to get drunk'
	c.	/ec-deng/	[ejjeng]	'to shoot'
	d.	/jo:c-deng/	[jo:jjeng]	'to carry on the head'

---

[9] The symbol /c/ represents the affricate /ts/ and the symbol /j/ represents the affricate /dz/. The stops /t/ and /d/ are dental. The symbol [ng] represents a velar nasal. All other symbols have their familiar interpretations. The consonantal phonemes of Pengo are p b v t d s z t d c j k g h m n n ng l r r y. The vowel phonemes are a e i o u. Vowel quantity is phonemic. (Vowel length is indicated in this paper by a colon after the vowel.) Burrow and Bhattacharya (1970) provide no comprehensive statement of phonotactics. Based on their grammar, however, the following phonotactic statements can be posited. Syllable onsets contain at most two consonants (though such syllables do not seem to be common). Syllable codas normally do not contain more than a single consonant. Verb roots can end in two consonants only if the first consonant is a coronal nasal and the second is a coronal stop. (Historically, verb roots could also end in /mb/ and /ng/; however, synchronically, it appears that the final obstruents of these two clusters are no longer part of the underlying representation since they do not appear in surface allomorphs.)

[10] As exemplified by the data in (13)-(15), Burrow & Bhattacharya (1970) note that some verb roots take *deng* as the infinitive suffix while other verb roots take *teng* as the infinitive suffix. It is not always predictable which of the two infinitive allomorphs a given verb root will take. As can be seen by comparing (13a-d) with (15a-d), some verb roots ending in /c/ take the allomorph *teng* while others take the allomorph *deng*. It is a lexical property of each of these roots as to whether they take *teng* or *deng*. This is also true of other verbal roots except for those ending in voiced obstruents. Roots ending in voiced obstruents always take the infinitive allomorph *deng* as evidenced by the data in (14). See Burrow & Bhattacharya (1970:95-97) for further discussion.

(16)	a.	/uj-t/	[ucc-]	'suck' (ptvs)
	b.	/hunj-t/	[hunc-]	'sleep' (ptvs)
	c.	/ne:nj-t/	[ne:nc-]	'breathe' (ptvs)
	d.	/vanj-t/	[vanc-]	'cook' (ptvs)

The data in (13)-(16) all show the application of the rule of Affricate Assimilation which applies to a suffix-initial (nonnasal) dental stop /t/ or /d/ when it immediately follows a root-final affricate /c/ or /j/. (There are no other affricates in Pengo.) The rule can be formalized as in (17) using the standard rule writing notation of Generative Phonolgy.

(17) **Affricate Assimilation**

```
   C    ---->  C       /    C     --
 +ant        +del rel     +del rel
 +cor        -ant
 -cont
 -son
```

The specific sequences that undergo the rule and the changes due to the rule's application are given in (18). (The output in parantheses in (18b) and (18c) is due to a later rule of Voicing Assimilation which is to be discussed shortly.)

(18) a. ct ---> cc
b. cd ---> cj (---> jj)
c. jt ---> jc (---> cc)
d. jd ---> jj

The forms in (15) and (16) provide evidence for an additional rule of Voicing Assimilation. This rule assimilates the voicing feature of the first obstruent in a cluster to that of the following obstruent. The rule is formalized below in (19) using the rule notation of Generative Phonology. The rule will be amended somewhat when further data are considered later.

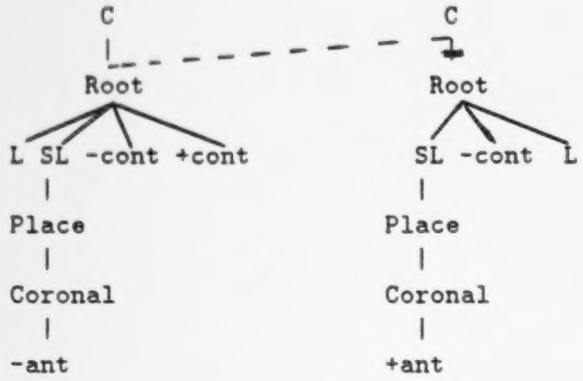
(19) **Voicing Assimilation**

```
[-sonorant] ---> [a voice] / __ -sonorant
                                 a voice
```

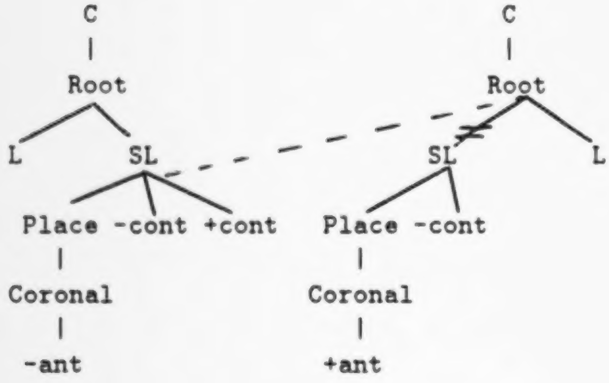
There is a further rule of Consonant Deletion that applies to the forms in (14b-d) and (16b-d) that will be discussed later.

The data in (13) and (14) are compatible with the feature [continuant] being located off of either the Root node (Theory R) or the Supralaryngeal node (Theory SL). Under Theory R, the affricate assimilation process shown in (13) and (14) would be a case of total assimilation (or Root node spreading); the Root node of the affricate spreads rightward while the original Root node of the /t/ delinks. This is shown in (20). Under Theory SL the affricate assimilations in (13) and (14) can be accounted for by the rightward spreading of the Supralaryngeal node of the affricate. This is seen in (21).[11]

(20)



(21)



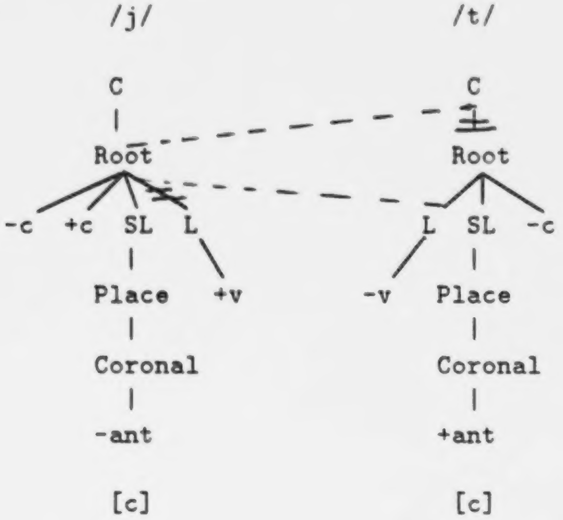
The data in (15) and (16a) show the application of the rule of Voicing Assimilation in addition to the application of the rule of Affricate Assimilation. The (regressive) Voicing

---

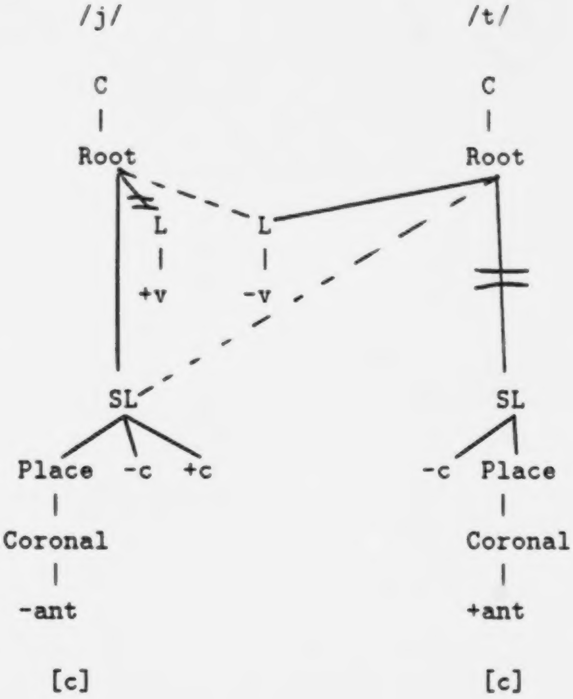
[11] Assuming the view of Sagey (1986b) that the [-continuant] [+continuant] features of an affricate are temporally ordered, affricate assimilation cannot simply involve the spreading of the [-continuant] [+continuant] features of the affricate onto the neighboring consonant; for doing such would result in crossing association lines. For an interpretation of what it means to cross association lines, see Sagey (1988).

Assimilation rule can simply be expressed (given the autosegmental formalism of Feature Geometry) as leftward spreading of the Laryngeal node of the second of two adjacent obstruents. This is compatible with the feature [continuant] either being located off of the Root node (shown in (22)) or off of the Supralaryngeal node (shown in (23)) in deriving these forms. (c=continuant)

(22)


/j/
/t/
C
C
Root
Root
-c
+c
SL
L
L
SL
-c
Place
+v
-v
Place
Coronal
Coronal
-ant
+ant
[c]
[c]

(23)


/j/
/t/
C
C
Root
Root
L
L
+v
-v
SL
SL
Place
-c
+c
-c
Place
Coronal
Coronal
-ant
+ant
[c]
[c]

At first glance the data in (16b-d) appear compatible with either Theory R or Theory SL. Under Theory R, a form like [hunc-] (in 16b) is derived from /hunj-t/ by three rules: Regressive Voicing Assimilation, Affricate Assimilation, and Consonant Deletion. The latter rule normally deletes a consonant that is both [+coronal] and [+distributed] in the environment after nasals and before /t/ or /d/. It is shown in (24).

(24) Consonant Deletion

```
   C ---> ∅ / C    --    C
 +cor       +nas       -cont
 +dist                 +ant
                       +cor
```

The data in (25) show that the rule of Consonant Deletion also applies to dental stops (in addition to the affricates in (14b-d) and in (16b-d)). The data in (26) show its lack of application to retroflex stops (which are [-distributed]). (In all the forms in (25) and (26) the past tense suffix /-t/ is added to a verb root to create a verb stem.)[12]

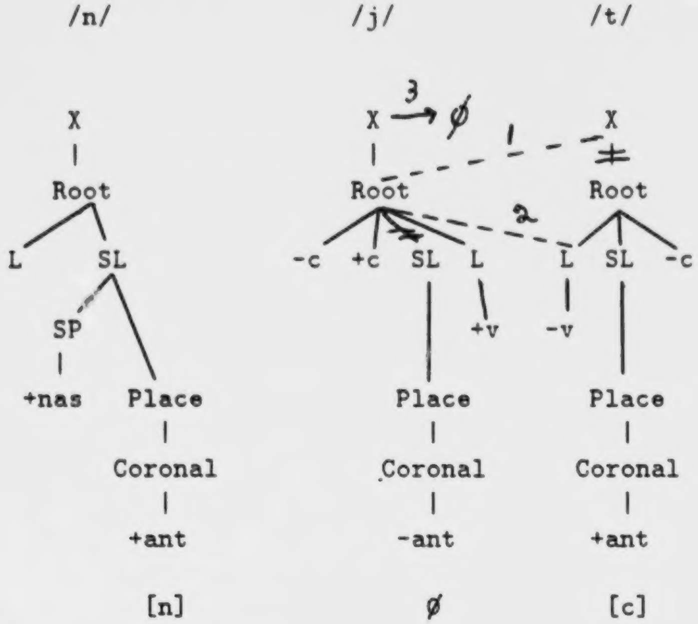
(25)	a.	/u:nd-t/	[u:nt-]	'take out'
	b.	/e:nd-t/	[e:nt-]	'dance'
	c.	/kund-t/	[kunt-]	'poke'
	d.	/nend-t/	[nent-]	'wipe'
(26)	a.	/anḍ-t/	[andt-]	'stick'
	b.	/kanḍ-t/	[kandt-]	'copulate'
	c.	/ge:nḍ-t/	[ge:nḍt-]	'learn'
	d.	/tinḍ-t/	[tinḍt-]	'sharpen'

Theory R seems able to account for the data in (16b-16d) by having all three rules apply to each. Affricate Assimilation and Voicing Assimilation would apply first (not crucially ordered) and then Consonant Deletion applies later. This is shown in (27) for the form [hunc-] (from /hunj-t/). (In this example, Affricate Assimilation, Voicing Assimilation, and Consonant Deletion are labeled 1-3, respectively)

---

[12] Burrow & Bhattacharya (1970) note that a nasal consonant is phonetically retroflexed before a retroflex stop. They do not say whether or not the final /t/ in the forms in (26) also become retroflexed after a retroflex stop.

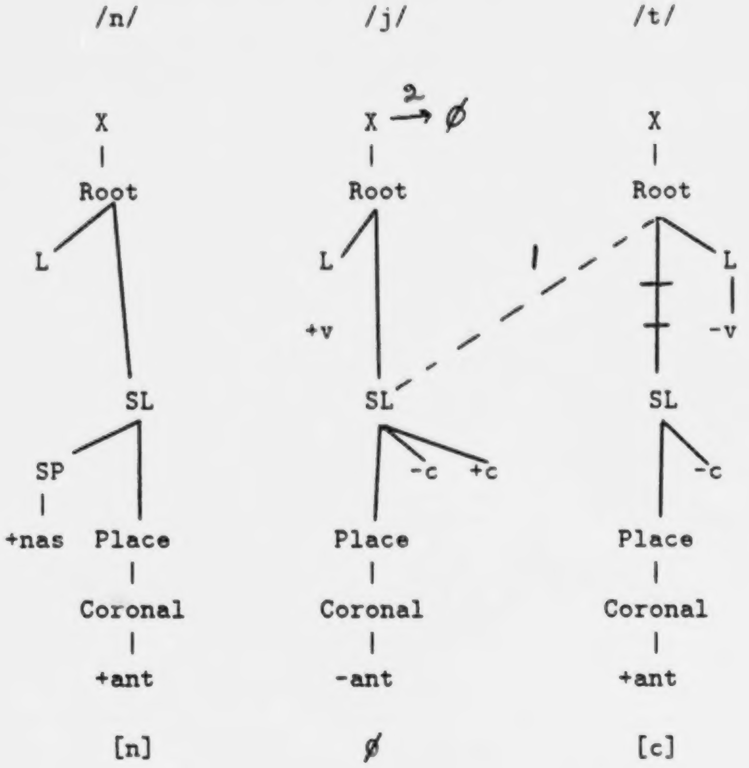
(27)


/n/
/j/
/t/
X
Root
L
SL
SP
+nas
Place
Coronal
+ant
X
3
ø
1
Root
2
-c
+c
SL
L
+v
Place
Coronal
-ant
X
Root
L
SL
-c
-v
Place
Coronal
+ant
[n]
ø
[c]

Notice that in (27) the rule of Voicing Assimilation (i.e., the leftward spreading of the Laryngeal node of the past tense suffix /t/) must apply, or else, the output after the application of Consonant Deletion would be the incorrect hunj-. Voicing Assimilation in (27,) above, has the effect of partially undoing the total assimilation brought about by the spreading of the Root node.

Theory SL is able to account for the data in (16b-d) as well. Here, however, only two rules are needed: Affricate Assimilation and Consonant Deletion. (To reiterate, Affricate Assimilation under Theory SL is an instance of the rightward spreading of the Supralaryngeal node of the affricate.) There is no need for Voicing Assimilation to apply. This is shown in (28) for the form [hunc-] from /hunj-t/ (where Affricate Assimilation and Consonant Deletion are labeled 1-2, respectively).

(28)



Both Theory R and Theory SL seem able to account for the data in (16b-d). The solution under Theory SL in (28) where the feature [continuant] is located off of the Supralaryngeal node can be considered superior because it only involves the application of two rules instead of three rules. However, as we will soon show, Theory R is actually unable to account for the data in (16b-d) because the rule of Voicing Assimilation which must apply to these data (as seen in the derivation of *hunc-* in (27)) is actually unable to apply. This is because Voicing Assimilation does not occur between adjacent obstruents immediately following a nasal. If Voicing Assimilation cannot apply to the forms in (16b-d), then the correct phonetic representations for these cannot be produced under Theory R (short of ad hoc stipulations).

We now discuss the rule of Voicing Assimilation (19) in more detail. Voicing Assimilation is unable to apply between two obstruents that immediately follow a nasal consonant, as is evident from the data in (26). Since Voicing Assimilation fails to apply in this environment, the derivations of the forms in (16b-d) under Theory R (as reflected in (27)) cannot be correct. This leads to the conclusion that Theory R, which posits that the feature [continuant] is located immediately under the Root node, cannot be maintained. In order to show that Voicing Assimilation fails to apply between adjacent obstruents immediately following a nasal we first must reconsider the data in (26). The data in (26) are meant to show that

Consonant Deletion does not apply to retroflex stops. Interestingly, in these forms, the rule of Voicing Assimilation fails to apply. Judging just from the data in (26), it might be surmised that Voicing Assimilation does not occur if a retroflex obstruent is involved. However, this is not the case as is made clear by the data in (29) in which the past tense marker /t/ is added to roots ending in a postvocalic /d/.

(29) a. /aḍ-t/ [aṭt-] 'be able'
b. /ka:ḍ-t/ [ka:ṭt-] 'burn'
c. /paḍ-t/ [paṭt-] 'be broke'
d. /naḍ-t/ [naṭt-] 'be severed'

The above data show that the condition that blocks Voicing Assimilation from applying to the forms in (26) cannot be the retroflex nature of the root-final stop; rather it is simply the presence of the preceding nasal consonant. That is, Voicing Assimilation fails to apply between two adjacent obstruents (as in (26)) when they are immediately precedeed by a nasal. Thus the Voicing Assimilation rule in (19) should be reformulated as in (30).

(30) Voicing Assimilation (revised)

```
     C   --->  [a voice] / [-nasal] __   C
   [-son]                               -son
                                        a voice
```

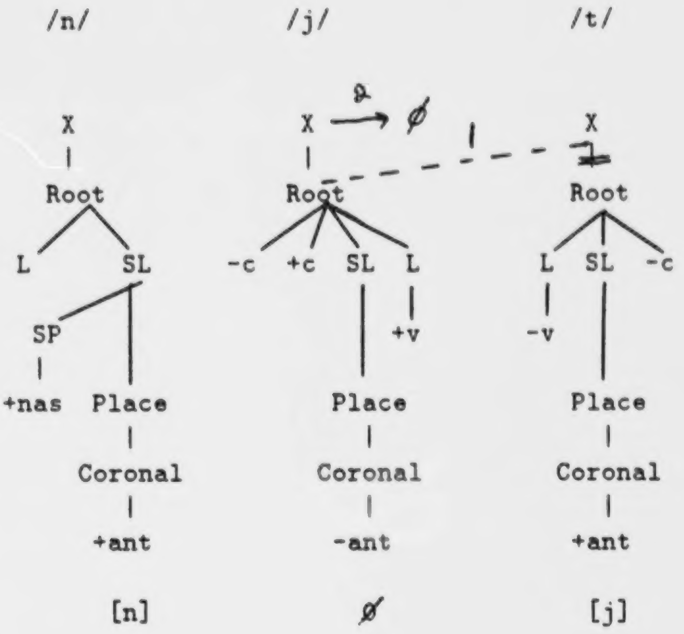
Independent evidence for the revised rule of Voicing Assimilation comes from the fact that some of the forms in (25) have alternant surface forms in which the post-nasal consonant is not deleted (i.e., Consonant Deletion fails to apply). These are shown in (31).

(31) a. /kund-t/ [kunt-] [kundt-] 'poke'
b. /nend-t/ [nent-] [nendt-] 'wipe'

As can be seen from the examples in (31), when /d/ fails to delete no regressive voicing assimilation occurs. The presence of the nasal consonant blocks the application of Voicing Assimilation from applying between the two following obstruents.

The condition on the rule of Voicing Assimilation, that Voicing Assimilation does not apply between two obstruents immediately following a nasal, means that the derivation of *hunc-* in (27) is not possible since it allows for Voicing Assimilation to apply in such an environment. In (32), the derivation of [hunc-] (from /hunj-t/) is shown as in (27) (in which the feature [continuant] is located off of the Root node) but without the application of Voicing Assimilation. (Affricate Assimilation and Consonant Deletion are labelled 1-2, respectively.)

(32)



Because the environment for Voicing Assimilation is not met in (32) it fails to apply; the rule of Affricate Assimilation (which, here, under Theory R, is expressed as spreading of the Root node since the feature [continuant] is located under the Root node) and the rule of Consonant Deletion apply producing the incorrect *hunj-*. On the other hand, in (28) where Theory SL is incorporated (i.e., where the feature [continuant] is located under the Supralaryngeal node), the rule of Affricate Assimilation and the rule of Consonant Deletion apply to produce the correct *hunc-*. Voicing Assimilation does not apply, and moreover, it could not apply since its environment is not met. We see, then, that the Pengo assimilation data is best analyzed under Theory SL.

Glottalization

In the previous subsection it was argued, based on a rule of Affricate Assimilation in the Dravidian language Pengo, that the feature [continuant] must be located under the Supralaryngeal node (i.e., that Theory SL is correct). In this subsection additional support for Theory SL is provided from the rules of supralaryngeal delinking in Kinyarwanda (a Bantu language of East Africa) and English. We first consider the Kinyarwanda evidence.

In Kinyarwanda a voiceless aspirated stop is pronounced as [h] when it follows a nasal consonant (with the nasal consonant subsequently devoicing). This is illustrated by the data in (33) taken from Sagey (1986b:35). (Although Sagey discusses the data in (33) she does not mention its possible relevance for the location of the feature [continuant].)

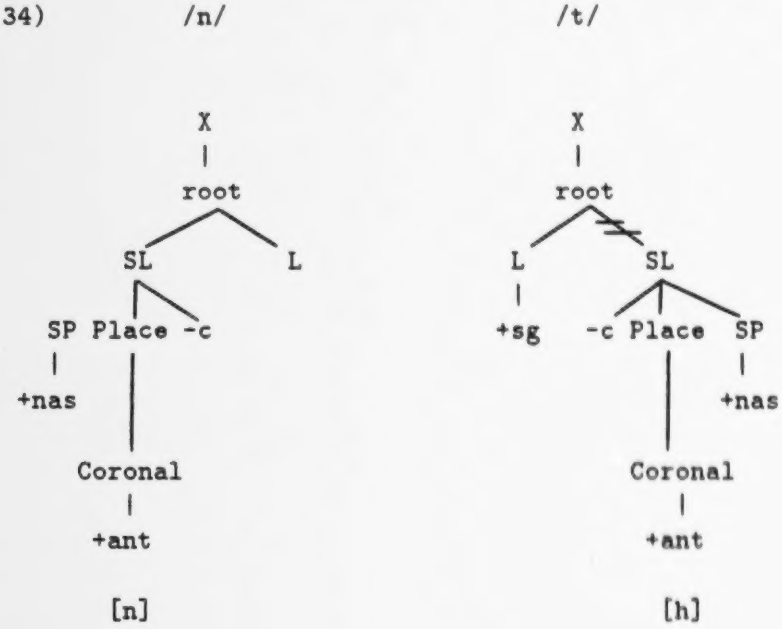
(33) a. /in-papuro/ [imhapuro] 'paper'

b. /n-toora/ [nhoora] 'I vote.'

c. /in-ka/ [i ha] 'cow'

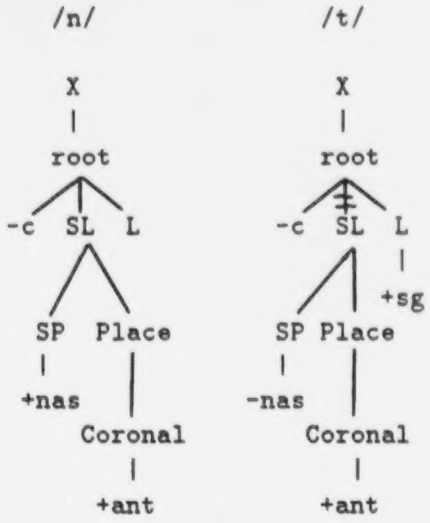
Given Theory SL, the change of an aspirated stop into the glottal [h] is simply accounted for by the delinking of the Supralaryngeal node. This is shown in (34) where /t/ becomes [h] after /n/. (The subsequent leftward spreading of the Laryngeal node is not shown.) (sg = spread glottis)

(34)



If, however, the feature [continuant] is located under the Root node, then, after the Supralaryngeal node of the aspirated stop delinks, the [-continuant] feature remains; the resulting segment would be both [-continuant] and [+spread glottis], but lacking any supralaryngeal features. This is shown in (35).

(35)



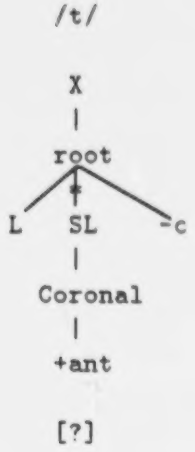
The resulting segment (on the right in (35)) is a laryngeal sound that is both [-continuant] and [+spread glottis]. Such a sound does not occur. A special additional rule would be needed to delete the [-continuant] feature. However, as shown in (34), no special rule is needed under theory SL where the resulting segment is an [h] and not a nonoccurring [-continuant, +spread glottis] laryngeal. Thus the Kinyarwanda rule of Supralaryngeal Delinking provides evidence for the correctness of Theory SL. For it is only Theory SL that correctly predicts the surfacing of an [h] when an aspirated stop loses its supralaryngeal features.

American English /t/-glottalization, at first glance, seems more compatible with Theory R than Theory SL. However, on closer examination, it is quite compatible with Theory SL. The rule of /t/-glottalization in one dialect of American English can be expressed by the following rule:

(36) t ---> ? / __ n

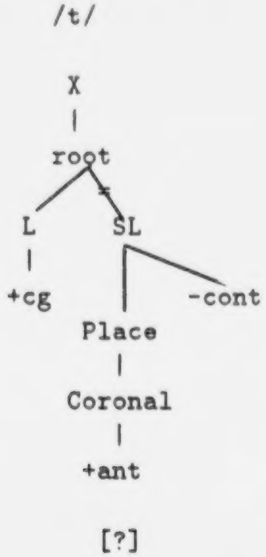
This rule changes /t/ into a glottal stop when before a syllabic /n/. Thus the underlying /t/ in such words as *fatten*, *cotton*, *kitten*, and *written* is pronounced as a glottal stop. As in Kinyarwanda, glottalization here can be understood as supralaryngeal delinking. That is, the phoneme /t/ loses all of its supralaryngeal features when before a syllabic [n]. Under Theory R it is not surprising that the supralaryngeal delinking of /t/ results in a glottal stop since the [-continuant] feature of /t/ does not delete. This is shown in (37).

(37)



Under Theory SL the resulting glottal stop is correctly predicted only if /t/ has the feature [+constricted glottis]. (In terms of features, glottal stops can be considered [+constricted glottis] or [-continuant] or both.) This is seen in (38).

(38)



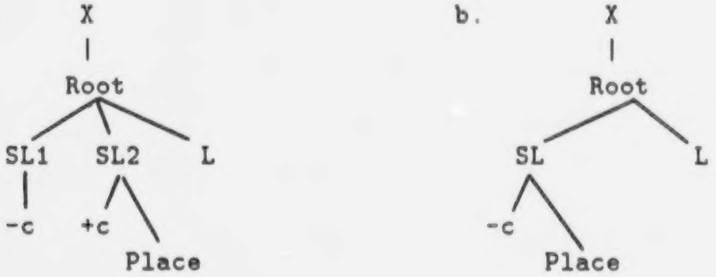
For Theory SL to account for the t/? alternation, the /t/ has to be [+constricted glottis] at the point where the glottalization rule (36) applies. It is thus suggested here that the segment that undergoes the rule in (36) is actually a preglottalized [?t]. The rule in (36) is related to the phenomena of preglottalization of English voiceless stops. Evidence that this is correct comes from the observation that /t/ alternates between being realized as a preglottalized stop (having the feature [+constricted glottis]) and being a full glottal stop in the phonetic realization of many morphemes (e.g., fat [fæ?t] vs. fatten [fæ?n̩] or write [ray?t] vs. written [rI?n̩]). Also, in some English dialects (such as Scots), a word-final voiceless stop becomes a glottal stop when after a vowel. This is the exact environment where there is preglottalization in most other English dialects. These facts suggest then that it is a preglottalized stop that loses its supralaryngeal features in such words as *fatten* and *written*. The resulting glottal stop is compatible with Theory SL.

In this subsection it has been shown that glottalization phenomena is best captured under Theory SL. Although the case of English glottalization is actually compatible with either Theory R or Theory SL, it is only the latter that adequatley accounts for the alternations between voiceless aspirated stops and [h] in Kinyarwanda. Thus glottalization phenomena support the representation of [continuant] as a supralaryngeal feature.

Deletion of [-continuant] in Basque

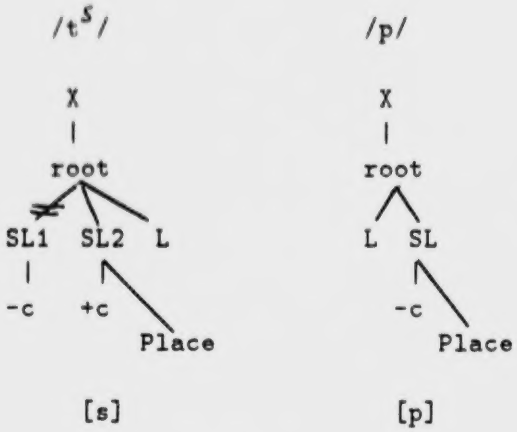
Further evidence that [continuant] is best represented as a supralaryngeal feature comes from the phenomena of stop deletion and affricate fricativization in Basque. This is discussed in detail by Hualde (1987) and the brief presentation here is based on Hualde's work. In Basque, a stop consonant deletes when it is before another stop consonant. Thus an underlying sequence like /kt/ is realized phonetically as just [t]. Moreover, if the two stop consonants have different features for voicing, the second stop will acquire the voicing value of the first. For example, an underlying sequence like /tg/ is realized phonetically as [k]. Also, in Basque, an affricate becomes a fricative when before another stop. Thus, an underlying sequence like /t^sp/ is realized phonetically as [sp]. Hualde notes that these two processes can be represented as a single process if Basque affricates are posited as having two supralaryngeal nodes and the feature [continuant] is represented as a supralaryngeal feature. Hualde's representation of a Basque affricate is shown below in (39a) while the representation for an ordinary stop is given in (39b). (Possible problems concerning the precise location of the Place node for affricates and whether there should be one or two of them is ignored.)

(39) a. b.

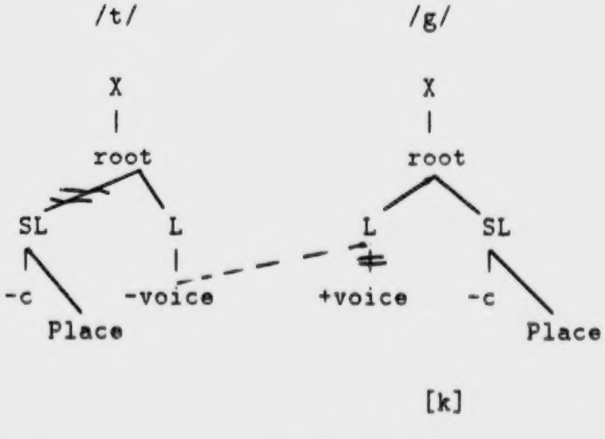


Given the representation of affricates and stops as in (39), Stop Deletion and Affricate Fricativization can be expressed as single process, namely the delinking (or deletion) of the (first) Supralaryngeal node. This is shown in (40) for the affricate-stop sequence /t^sp/ which is realized phonetically as [sp] and in (41) for the stop-stop sequence /tg/ which is realized phonetically as [k].

(40)

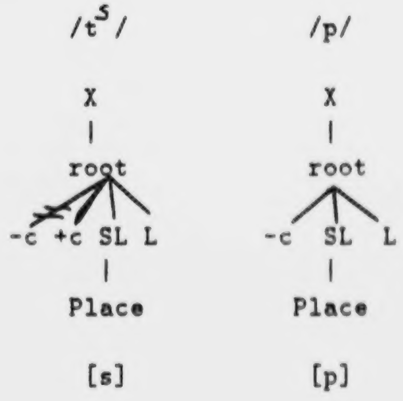

/t^s/
X
root
SL1
SL2
L
-c
+c
Place
[s]
/p/
X
root
L
SL
-c
Place
[p]

(41)


/t/
X
root
SL
L
-c
Place
-voice
/g/
X
root
L
SL
+voice
-c
Place
[k]

If the feature [continuant] were located immediately under the Root node and not under the Supralaryngeal node then Affricate Fricativization (40) and Stop Deletion (41) would involve different delinkings. Affricate Fricativization would involve delinking of the [-cont] feature, shown in (42), while Stop Deletion would involve the delinking of the entire Root node.

(42)



Thus, in order to capture Basque Affricate Fricativization and Stop Deletion as instances of the same process (i.e., delinking of a supralaryngeal node), the feature [continuant] must be considered a supralaryngeal feature.

It should be pointed out here that Hualde's representation of affricates (39a) in which there are two supralaryngeal nodes appears incompatible with the analysis of Pengo affricates (e.g., in (21)) offered earlier in this paper. This is because Hualde's representation of affricates, applied to Pengo, would not permit Affricate Assimilation to occur by means of supralaryngeal node spreading. It could only be expressed as an instance of root node spreading. However, as we have seen in (32), this leads to deriving incorrect forms for data like that in (16b-d). In order to rectify this situation two possibilities are suggested. First, certain details of Feature Geometry may be language-specific. While the general division into laryngeal and supralaryngeal features, for example, may be universal other aspects may be language-specific. Thus the exact representation of an affricate may depend on the language in question. For example, Pengo affricates might only have one Supralaryngeal node while Basque affricates have two Supralaryngeal nodes. In both languages, though, [continuant] would be a supralaryngeal feature. Alternatively, it is possible that the Basque Affricate Fricativization and Stop Deletion are different rules despite having the same conditioning environment. If such is the case, then the Basque data would really not have direct bearing on whether or not [continuant] should be considered a supralaryngeal feature.

Conclusion

In this paper it has been argued based on a variety of phenomena in diverse languages that [continuant] is a supralaryngeal feature. Our argumentation though is really not complete. We have argued that [continuant] is a supralaryngeal feature but we have not argued where under the Supralaryngeal node it is to be located. Any of the geometries outlined in (7)-(9) is compatible with the proposal. Moreover, there have been other proposals about the feature [continuant] that we have not discussed in this paper. For example, Iverson (1988) argues based on Korean data that [continuant] is best represented as being directly under the Root node. While the Korean data he considers is not incompatible with [continuant] being under the Supralaryngeal node, the data does seem better analyzed in some other way. It is quite possible that when more languages are analyzed phonologically it may be the case that it is a language-specifc matter whether or not [continuant] is a supralaryngeal feature. We leave this matter for future research.

References

Ali, L., Daniloff, R., & Hammarberg, R. (1979). Intrusive stops in nasal-fricative clusters: An aerodynamic and acoustic investigation. *Phonetica*, **36**, 85-97.

Anderson, S., (1976). Nasal consonants and the internal structure of segments. *Language*, **52**, 326-344.

Archangeli, D. & Pulleyblank, D. (1986). *The content and structure of phonological representations*. Unpublished ms., University of Arizona, Tucson, Arizona, and University of Southern California, Los Angeles, California.

Burrow, T. & Bhattacharya, S. (1970). *The Pengo language*. Oxford: Clarendon Press.

Cho, Y. (1988). Korean assimilation and Feature Geometry. Paper presented at the Seventh West Coast Conference on Formal Linguistics, University of California, Irvine, February 1988.

Clements, G. (1985). The geometry of phonological features. *Phonology Yearbook*, **2**, 223-252.

Clements, G. (1987). Phonological feature representation and the description of intrusive stops. In Bosch, A., Need, B., & Schiller, E. (Eds.), *Papers from the 23rd regional meeting of the chicago linguistic society-part two: Parasession on autosegmental and metrical phonology*, 29-50. Chicago: Chicago Linguistic Society.

Clements, G. & Keyser, J. (1983). *CV phonology*. Cambridge: MIT Press.

Fourakis, M. and Port, R. (1986). Stop epenthesis in English. *Research in phonetics and computational linguistics, report no. 5*, 37-71. Bloomington, IN: Departments of Linguistics and Computer Science, Indiana University.

Halle, M. (1986). *On speech sounds and their immanent structure*. Unpublished ms., MIT, Cambridge, Massachusetts.

Hayes, B. (1986) Assimilation as spreading in Toba Batak. *Linguistic Inquiry*, **17**, 467-499.

Hualde, J. (1987). On Basque affricates. In Crowhurst M. (Ed.), *Proceedings of the sixth west coast conference on formal linguistics*, 77-89. Stanford CA.: The Stanford Linguistics Association.

Iverson, G. (1988). On the category supralaryngeal. Paper presented at the Annual Meeting of the Linguistic Society of America, New Orleans, Dec. 27-30.

Levin, J. (1985). *A metrical theory of syllabicity*. Unpublished doctoral dissertation, MIT, Cambridge, Massachusetts.

Sagey, E. (1986a). On the representation of complex segments and their formation in Kinyarwanda. In Wetzels, L. & Sezer, E. (Eds.), *Studies in compensatory lengthening*, 251-295. Dordrecht: Foris.

Sagey, E. (1986b). *The representation of features and relations in non-linear phonology*. Unpublished doctoral dissertation, MIT, Cambridge, Massachusetts.

Sagey, E. (1988). On the ill-firmedness of crossing association lines. *Linguistic Inquiry*, **19**, 109-118.

Training Japanese Listeners to Identify /r/ and /l/: A First Report[1]

John S. Logan, Scott E. Lively, and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This research was supported, in part, by NIH Research Grant *NS-12179-12* to Indiana University in Bloomington, IN, and, in part, by a Marie Z. Sinnott Fellowship from Indiana University to the first author. The authors thank Natsuko Tsujimura and N. John Castellan for their assistance in obtaining subjects.

Abstract

Native speakers of Japanese learning English generally have difficulty differentiating the phonemes /r/ and /l/, even after years of experience with English. Previous research that attempted to train Japanese listeners to distinguish this contrast using synthetic stimuli showed little success, especially when transfer to natural tokens containing /r/ and /l/ was tested. In the present study, a procedure that differed from these earlier attempts was used. Japanese subjects were trained in a minimal pair identification paradigm using as stimuli multiple natural exemplars contrasting /r/ and /l/ from a variety of phonetic environments. A pretest-posttest design containing natural tokens was used to assess the effect of training. Analysis of data from six subjects showed that the new procedure was more robust than earlier training techniques. Small but reliable differences in performance were obtained between pretest and posttest scores. The results demonstrate the importance of stimulus variability and task-related factors in training second language learners to perceive novel phonetic contrasts that are not distinctive in their native language.

Training Japanese Listeners to Identify /r/ and /l/: A First Report

When listeners are presented with speech stimuli from phonetic categories that are not used in their own language, they typically show performance that is not as good as a native speaker of the language from which the phonemes were obtained. A developmental account of this phenomenon suggests that within the first year of life the infant begins to move from language-universal abilities to the language-specific abilities that are characteristic of the adult (e.g., Werker, in press). Language-universal refers to how infants are able to discriminate virtually any phonetic contrast used in a language, regardless of the environment in which they are raised, while language-specific refers to the much more restricted abilities of adults to discriminate or identify stimuli from phonetic categories not used in their native langauge. The transition from language-universal to language-specific abilities is a product of the interaction between innate perceptual mechanisms and early linguistic experience (Aslin & Pisoni, 1980). Early experience serves to modify the sensitivity of the child's perceptual system such that only those phonetic contrasts that denote differences in meaning remain distinctive. Learning to perceive a nonnative phonetic contrast thus generally poses special problems for adults.

Much of the work in cross-language speech perception is based on findings obtained from experiments investigating the phenomenon of categorical perception. One characteristic of categorical perception is that listeners can discriminate stimuli drawn from different phonetic categories but not from within the same phonetic category (Liberman, Harris, Hoffman, & Griffith, 1957). However, under certain conditions, within-category discrimination is possible. Factors that facilitate the differentiation of within-category stimuli have generally relied on accessing sensory memory, using minimal uncertainty procedures, and extensive training of listeners. Some of these factors, as well as their application to the perception of nonnative phonetic contrasts, are described briefly below.

Pisoni (1973) demonstrated the role of sensory memory by showing that if the interval between presentation of stimuli in an AX 'same' – 'different' discrimination task was sufficiently short, listeners displayed improved discrimination of synthesized vowel stimuli from within a phonetic category compared to performance at longer intervals. Werker and Tees (1984) employed the procedure developed by Pisoni using natural CV stimuli from a place of articulation contrast occuring in Hindi. They found that native speakers of English could discriminate these nonnative stimuli more accurately if the interval between the stimuli was reduced.

Additional evidence for how sensory memory could be utilized to improve within-category discrimination performance was provided by Pisoni and Lazarus (1974). Subjects were presented stimuli varying in voice-onset time (VOT) with endpoints corresponding to /ba/ and /pa/. Pisoni and Lazarus compared performance using two different discrimination tasks, an ABX task and a 4IAX task, preceded by two types of identification tasks. In a 4IAX discrimination task subjects are presented two pairs of stimuli, one containing two identical stimuli and the other containing two different stimuli; subjects are required to indicate which pair contains the 'same' or 'different' stimuli. In an ABX task, subjects are required to indicate if a third stimulus is the same as the first (A) or the second (B) stimulus. In one of the identification tasks, subjects were presented the stimuli in random order and asked to label them using /ba/ and /pa/ labels (Random presentation), while in the other identification task, subjects simply heard the stimuli presented in ascending and descending order according to increasing and decreasing VOT values (Sequential presentation). Pisoni and Lazarus found that within-category discrimination was best in the 4IAX task preceded by the Sequential presentation of stimuli. They argued that using the 4IAX task, coupled with the Sequential identification task, permitted subjects to access sensory memory, enabling them to compare stimuli based on acoustic differences rather than on more abstract phonetic codes.

Carney, Widin, and Viemeister (1977) demonstrated the effects of task uncertainty and training in an experiment designed to explore the ability of subjects to discriminate stimuli from within a phonetic category. They used a continuum of synthesized stimuli varying in voice-onset time (VOT) that is typically divided into two phonetic categories by native speakers of English. Carney et al. used a fixed-standard AX discrimination procedure in which the first stimulus of an AX pair is always the same for a block of trials and only the second stimulus of the pair varies. The fixed-standard procedure is a low-uncertainty task used in auditory psychophysics that enables listeners to detect small acoustic differences between stimuli. In addition, Carney et al. used subjects that were highly experienced with psychophysical procedures and who received extensive training with the task. Carney et al. found that using such procedures, listeners could discriminate stimuli from any location along the stimulus continuum. Although they did not use stimuli from a nonnative phonetic category, their results indicated that, using the appropriate methodology, even very small differences among speech stimuli could be discriminated. Later work which attempted to train listeners to discriminate nonnative phonetic contrasts was heavily influenced by the success of Carney et al.'s procedure.

A final demonstration of how task factors can affect the identification and discrimination of nonnative stimuli comes from an experiment that was carried out by Pisoni, Aslin, Perey, and Hennessy (1982). They found that simply by providing the appropriate number of response categories, native speakers of English could easily learn to label stimuli from a VOT continuum into three categories that earlier work indicated could only be divided into two categories. Pisoni et al.'s results suggested that simple laboratory training procedures could be used to modify the phonetic categories of adult listeners presented with stimuli from nonnative phonetic categories varying along the VOT dimension.

Taken together, the work described above demonstrates that the nature of the task plays an important role in determining how well listeners are able to discriminate nonnative speech contrasts. One factor that these studies did not address was the degree to which the changes in perception demonstrated in laboratory procedures generalize to more "ecologically valid" situations outside the laboratory. Even if listeners can perform superhuman feats of discrimination in a laboratory setting, the practical goal of perceiving nonnative contrasts in more realistic contexts may not be attained. Tasks utilizing auditory sensory memory and minimal uncertainty procedures work because they enable a listener to take advantage of very small differences in the acoustic information that signal a phonetic category for the native listener but are not salient for the nonnative listener under most non-laboratory conditions. Thus, there exists a need for procedures which can be used by listeners to improve their perception of nonnative phonetic contrasts that are likely to be useful to the individual once he or she has left the laboratory.

In 1984, Strange and Dittman reported the results of an experiment in which they attempted to modify the perception of /r/ and /l/ by Japanese listeners. Strange and Dittman's work differed from the studies described above because not only did they seek to demonstrate the capabilities of the perceptual system but they also wanted to see if what subjects learned in their experimental task could be generalized to other contexts.

Before describing Strange and Dittman's procedure in detail, it is useful to consider what is known about the perception of /r/ and /l/ by Japanese listeners. The difficulty that Japanese listeners have with the /r/ - /l/ contrast is well documented. Even after years of living in an English speaking environment, the performance of Japanese listeners presented synthesized stimuli contrasting /r/ and /l/ in identification and discrimination tasks differs from the performance of native speakers of English when tested in the same tasks (MacKain, Best, & Strange, 1981).

The difference in performance between inexperienced Japanese subjects and native speakers of English is even larger; inexperienced Japanese generally have more poorly defined category boundaries and their discrimination functions are typically flat and close to chance performance (Miyawaki, Strange, Verbrugge, Liberman, Jenkins, & Fujimura, 1975; MacKain, Best, & Strange, 1981; Mochizuki, 1981). In addition, natural speech contrasting /r/ and /l/ is also poorly perceived by Japanese listeners, whether it is produced by a native English speaker or a native Japanese speaker (Goto, 1971; Sheldon & Strange, 1982). Several studies have shown that the performance of Japanese subjects presented with /r/ and /l/ is not uniformly poor but instead is dependent on the context within a word in which /r/ and /l/ are located (Gillette, 1980; Mochizuki, 1981; Sheldon & Strange, 1982). In general, performance was worst for /r/ and /l/ in initial positions in either singleton and consonant cluster environments and intervocalic positions, while performance was relatively good for /r/ and /l/ in final position, again in either singleton and cluster environments. According to Sheldon and Strange (1982), while there is no obvious a priori phonological explanation for this context effect, preliminary acoustic analyses of stimuli containing /r/ and /l/ in these phonetic contexts suggests that there are systematic acoustic differences among the different phonetic environments.

The goal of Strange and Dittman (1984) was to improve the perception of /r/ and /l/ by Japanese listeners. Using the same AX discrimination task employed by Carney et al. (1977), they trained Japanese subjects to discriminate stimuli from a synthesized 'rock'-'lock' continuum for 12 to 14 sessions. The effectiveness of the training procedure was evaluated using a pretest-posttest design in which initial levels of performance for natural speech were compared with performance after discrimination training on these synthetic tokens. Strange and Dittman found that although performance improved during training, the effect of training did not generalize to the stimuli used in the posttest. They concluded that the laboratory training they used was of limited use in altering the perception of /r/ and /l/.

Task-related factors may have been responsible for the failure of Strange and Dittman (1984) to find generalization. The logic of their methodology was that if subjects could learn to focus their attention on the physical cues that distinguish /r/ and /l/ in a low uncertainty task, what they learned would also be usable under more demanding conditions. However, most of the procedures designed to show that adult listeners can perceive nonnative contrasts are fragile in that they enable access to perceptual mechanisms that in most situations are not directly accessible. In general, most models of speech perception posit a mechanism that transforms an initial veridical representation of the speech signal into an abstract phonetic represention (Pisoni & Luce, 1987). Listeners usually have access to only the product of this process, the phonetic representation. Thus, training which capitalizes on utilizing information contained in sensory memory has a high probability of not generalizing to more 'natural' conditions unless some more permanent memory code can be developed.

The goal of the present investigation was see under what conditions a group of native Japanese speakers could learn to identify the phonemes /r/ and /l/. More importantly, we wanted to develop a procedure that would likely prove to be useful in settings outside the laboratory. The present experiment was motivated, in part, by the earlier work of Strange and Dittman (1984). We employed the same pretest-posttest design Strange and Dittman used to evaluate the effectiveness of training on the perception of /r/ and /l/ by a group of Japanese listeners. However, our training methods differed in several important ways from theirs.

The first major change in the training procedure involved choosing a different task, one that would be less 'fragile' and more apt to result in transfer to other contexts. The procedure we chose was a two-alternative forced choice identification task, primarily because it would be more similar to what listeners would be required to do in settings outside the laboratory than the AX discrimination task employed by Strange and Dittman.

The other major change involved the choice of stimuli used during training. First, we used natural tokens instead of synthesized stimuli. Second, the set of stimuli contrasted

/r/ and /l/ in a wide variety of phonetic contexts instead of only one phonetic context. And third, the stimuli were produced by multiple talkers instead of only one (synthetic) talker. The rationale for the differences between stimuli used in the present experiment and in Strange and Dittman's study involved the issue of stimulus variability. In a well-known study, Posner and Keele (1968) showed that subjects trained to sort visual stimuli into categories performed better when given previously unseen stimuli if the stimuli they were trained on had a great deal of variability than if the stimuli had very little variability. Thus, one possible reason for why Strange and Dittman failed to find transfer of what was learned during training to novel stimuli was that their subjects were trained using stimuli that were not sufficiently variable to form 'robust' phonetic categories for /r/ and /l/. The role of stimulus variability in creating an adequate representation of /r/ and /l/ was acknowledged by Strange and Dittman (1984) themselves when they suggested that future work "include training of the contrast with more than one set of stimuli and in more than one phonetic context." Adding talker variability was a further attempt to promote the formation of robust phonetic categories that would be stable under a wide variety of conditions.

Finally, in addition to assessing the effectiveness of training in a posttest, we further tested generalization by presenting subjects with novel words produced by both new and old talkers. Recently, Mullennix, Pisoni, and Martin (in press) have shown that talker variability has important perceptual consequences for the speed and accuracy of processing spoken words. In light of Mullennix et al.'s findings, we decided that it would be useful to know if the ability of listeners to generalize to novel stimuli depended on the characteristics of the talkers used during training. We assumed that the variability of the talkers used during training would be sufficient to overcome talker-specific learning. That is, after training, subjects should be able to correctly identify novel words containing /r/ and /l/ produced by a novel talker. Such a test could be considered to be the most stringent measure for determining if what the Japanese listeners had learned during training would generalize to other contexts.

Method

Subjects. Subjects in the present experiment were six native speakers of Japanese living in the Bloomington, Indiana area. All were students at Indiana University and had lived in the U.S. for periods ranging from six months to three years at the time of testing. All the subjects reported that they considered their proficiency with spoken English to be less than their ability to deal with written English. No subjects reported any history of a speech or hearing disorder. Subjects were paid $5.00 for each session.

Stimuli. A computerized database containing approximately 20,000 words (*Webster's Seventh Collegiate Dictionary*, 1967) was searched to locate all minimal pairs contrasting /r/ and /l/. A total of 207 minimal pairs were found. These words contrasted /r/ and /l/ in word-initial and final positions, in singleton and cluster environments, and in intervocalic positions. Six talkers, four male and two female, recorded the words in an IAC sound-attentuated booth using an Electro-Voice D054 microphone. Talkers were given no special instructions concerning pronunciation of the words, which were presented in a random order on a CRT monitor inside the recording booth. The words were low-pass filtered at 4.8 kHz and digitized at 10 kHz using a 12-bit analog-to-digital converter. The digitized waveform files were edited and then equated for RMS amplitude.

The stimuli were pretested with native speakers of English to ensure their intelligibility. After pretesting, a set of 136 stimuli (68 minimal pairs) from five talkers was selected for use in the training phase of the experiment. A set of 96 additional stimuli from a new talker were selected for use in a test of generalization. A final set of 98 additional stimuli from a talker who was heard during training was selected for use in a second test of generalization. In addition, the 32 pretest-posttest words used by Strange and Dittman (1984) were recorded by a male talker not included in either the training and generalization stimulus sets. These stimuli were processed in the same way as the other stimuli used in the present experiment.

Procedure. The experimental design employed a pretest-posttest procedure that was closely modeled after Strange and Dittman (1984). Before training began, subjects were presented 16 minimal pairs contrasting /r/ and /l/, each presented twice (Pretest). They were required to identify the word presented from a minimal pair printed in an answer booklet by making a written response in the booklet. The same test was administered again after training (Posttest). The words used in the pretest and posttest were the same as those used by Strange and Dittman (1984). The pretest-posttest required approximately 20 minutes to complete. The pretest was administered twice prior to training for three of the subjects in order to assess the extent to which mere exposure to the words used in the pretest might contribute to any improvements in performance observed in the posttest. A two-week period elapsed between the administration of the first and second pretest during which the subjects did not participate in the experiment.

The training phase also used an identification task. Subjects were presented a word from a minimal pair contrasting /r/ and /l/. They were required to identify the stimulus presented from a minimal pair presented on a CRT screen by pressing a button on a response box. Feedback was given during the training task. If the subject made a correct response, the next trial began. If the subject made an incorrect response, the minimal pair remained on the CRT screen and a light on the response box corresponding to the correct response was illuminated followed by a second presentation of the stimulus, afterwhich the next trial began. Stimuli from a set of 68 minimal pairs were each presented twice during a session, yielding a total of 272 trials in each session. During each training session, stimuli from only one talker were presented. Subjects cycled through the set of five talkers used during training three times (always in the same order) for a total of fifteen training sessions. Subjects were tested individually during training in a session that lasted approximately 40 minutes.

After the posttest, three of the subjects were tested to assess the degree to which training generalized to novel stimuli. The first test of generalization (TG1) consisted of 96 novel words from minimal pairs contrasting /r/ and /l/ produced by a new talker. A second test of generalization (TG2) consisted of 98 novel words from minimal pairs contrasting /r/ and /l/ produced by Talker 4 who subjects had heard during training. These were all new words that the subjects had not heard before. In both tests of generalization, the task was identical to that used during training except that subjects did not receive any feedback. The tests of generalization were administered individually.

Subjects were tested in a quiet sound-treated room containing individual cubicles. Each cubicle was equipped with a desk, a two-button response box, and a CRT monitor. Stimuli were presented over matched and calibrated TDH-39 headphones at 80 dB SPL. Presentation of stimuli and collection of responses was under the control of a laboratory computer (PDP-11/34). During training and tests of generalization, both identification responses and latencies were collected. Latencies were measured from the onset of the stimulus presentation.

Results

Pretest-Posttest

Results from the pretest-posttest phase of the experiment will be described first. The subset of subjects who were given the pretest twice prior to training showed no improvement in performance from the first administration ($M = 77.03$ correct) to the second administration ($M = 76.53$ correct), $t(2) = -0.83$. Thus, mere repetition of the vocabulary provided no measurable improvement in identification performance. No significant difference in pretest performance was found between the group of subjects administered the pretest twice (M pretest 1 and pretest 2 = 76.78) and the subjects administered the pretest only once ($M = 80.2$), $t(5) = 0.353$. In all subsequent analyses, the data for the two groups were combined.

Insert Figure 1 about here

The percentage of correct responses in the pretest and posttest is plotted in Figure 1. Overall, there was a significant increase in the percentage of correct responses from the pretest ($M = 78.49$) to the posttest ($M = 85.42$), $F(1, 5) = 38.47$, $p < .005$. We conclude from this result that subjects were, in fact, able to transfer what they learned about /r/ and /l/ during training to the posttest stimuli. This result attests to the efficacy of the training procedure used in the present experiment, and contrasts markedly with Strange and Dittman's apparent failure to obtain any differences in performance between pretest and posttest.

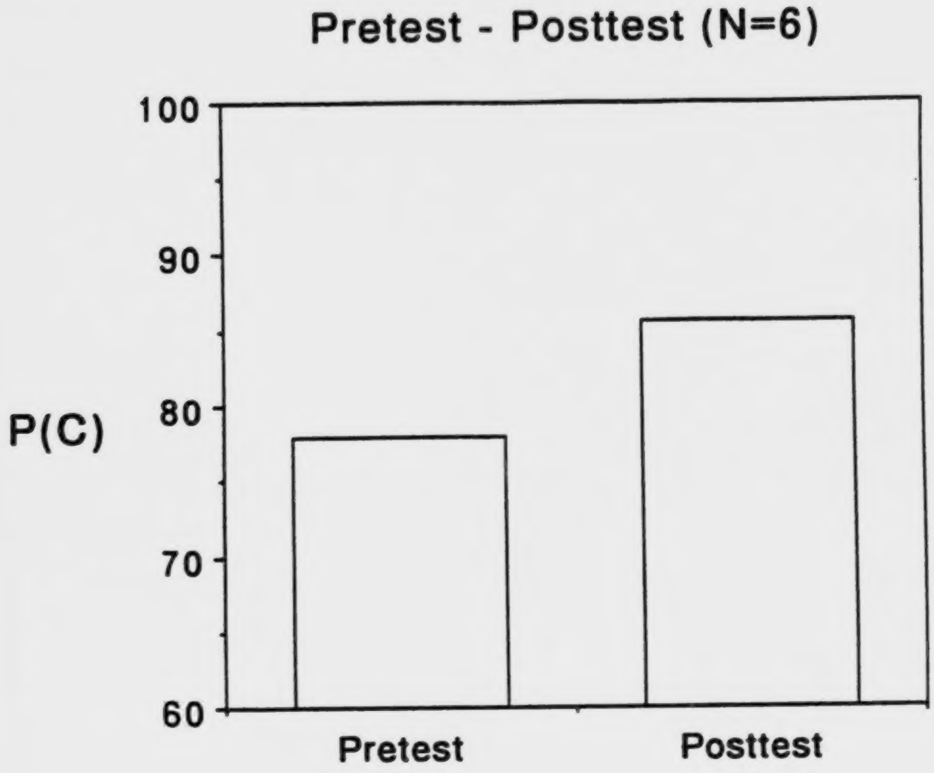


Figure 1. Mean percentage of correct responses in the pretest and posttest.

Insert Figure 2 about here

The percentage of correct responses for each of the four phonetic environments in the pretest and posttest is plotted in Figure 2. An analysis of the four phonetic environments showed significantly better performance for words contrasting /r/ and /l/ in final ($M = 96.87$) and intervocalic ($M = 83.07$) positions than for words contrasting /r/ and /l/ in initial position (singleton [$M = 79.95$] and initial clusters [$M = 68.23$]), $F(3, 15) = 6.32$, $p < .01$. Moreover, there was a significant interaction between pretest-posttest performance and phonetic environments, $F(3, 15) = 3.1$, $p < .05$. Performance for words from initial clusters and intervocalic environments improved markedly from the pretest to the posttest. In contrast, performance for words from the other two environments only improved slightly from the pretest to the posttest.

Insert Figure 3 about here

Training

The results from the training phase of the experiment are described next. An analysis of variance comparing week (weeks 1-3), talker (talkers 1-5) and phonetic environment (environments 1-5) was carried out. Figure 3 shows the percentage of correct responses as a function of week. A significant effect of week was obtained, $F(2, 8) = 14.85$, $p < .01$. Identification accuracy improved significantly from week 1 to week 2 but the improvement from week 2 to week 3 was not statistically reliable. Despite the highly variable stimulus set, modification of the subject's perceptual mechanisms appeared to take place rapidly within the first two weeks of training.

Insert Figure 4 about here

Figure 4 shows the percentage of correct responses as a function of the five talkers who produced the stimuli. An examination of Figure 3 indicates substantial differences in performance among talkers. In the ANOVA, a significant effect of talker was obtained, $F(4, 16) = 21.88$, $p < .0001$, confirming the trends shown in Figure 3. Talker 4 and Talker 5 were significantly more intelligible than Talkers 1, 2 and 3. Even though all the stimuli were tested with native speakers of English to ensure high intelligibility, the Japanese listeners appeared to be much more sensitive than the English listeners to individual differences among talkers. Until acoustic analyses have been completed, the precise source of this variation remains unknown. Analyses are currently underway.

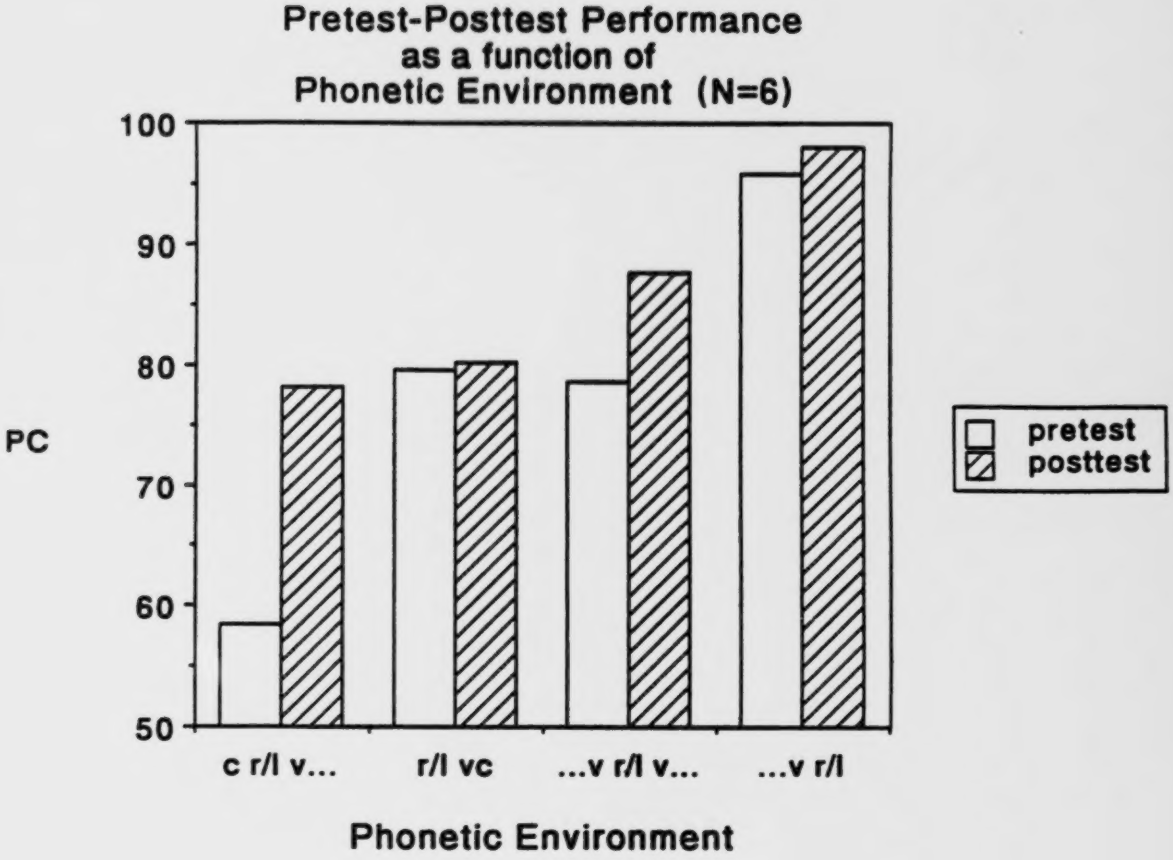


Figure 2. Mean percentage of correct responses in the pretest and posttest as a function of phonetic environment.

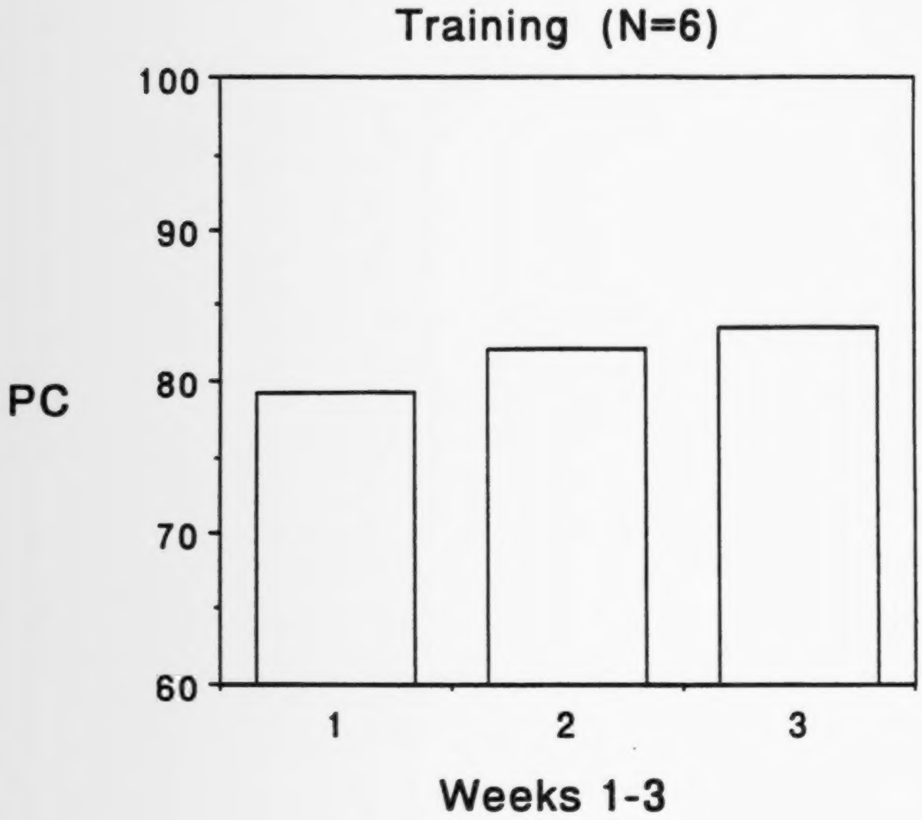


Figure 3. Mean percentage of correct responses during training as a function of week.

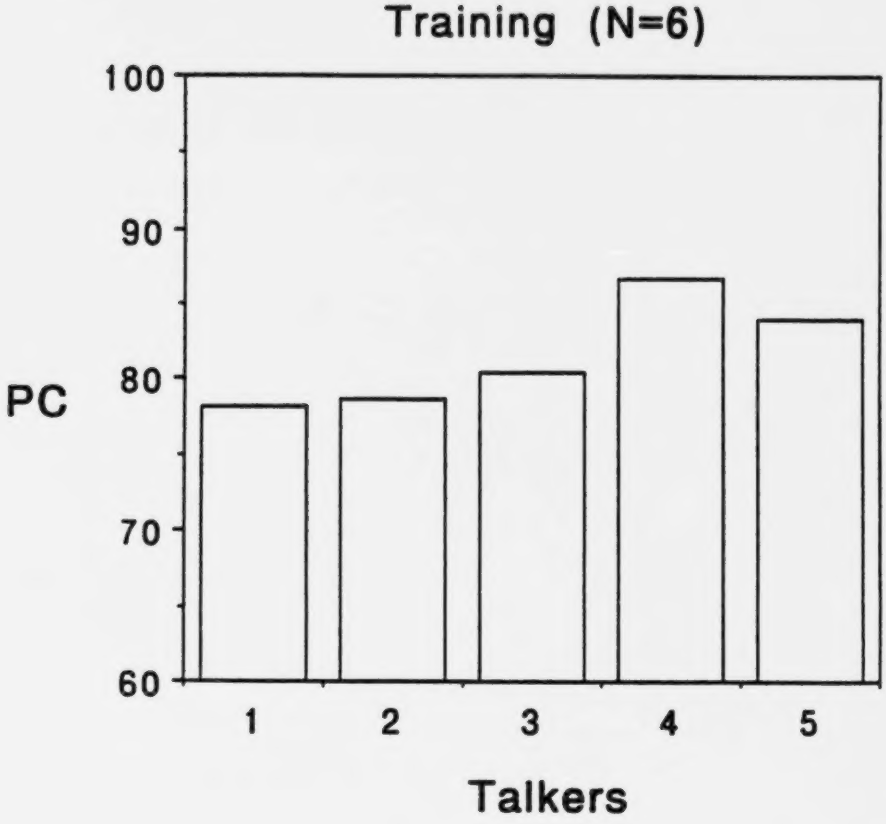


Figure 4. Mean percentage of correct responses in training as a function of talker.

Insert Figure 5 about here

Figure 5 shows the percentage of correct responses as a function of phonetic environment. For two of the environments, final singleton and clusters, performance is close to asymptotic whereas performance in the remaining environments ranges between 70% and 80% correct. There was also a significant effect of phonetic environment, $F(4, 16) = 16.96$, $p < .001$. Performance during training was best in final singleton and cluster position. Performance was significantly lower in initial singleton and initial cluster position, as well as in intervocalic positions. Thus, the effect of different phonetic environments found in the pretest-posttest data was obtained in training as well, replicating once more the earlier findings of Mochizuki (1981) and others.

Insert Figure 6 about here

Figure 6 shows the percentage of correct responses for each talker as a function of phonetic environment. For initial singleton and cluster environments, performance was uniformly good for all talkers whereas in the other environments, performance was consistently lower and varied widely as a function of talker. The interaction between talker and phonetic environment was significant, $F(16, 64) = 3.01$, $p < .001$. This result indicates that some talkers were much better than others in producing /ɹ/'s and /l/'s in environments that, in general, are poorly perceived. In the remaining phonetic environments, talker variability apparently made little difference in performance.

Insert Figure 7 about here

Response times were also collected during training. An ANOVA comparing the mean response times for correct responses across week (weeks 1-3), talker (talkers 1-5), and phonetic environment (environments 1-5) was carried out. A significant effect of talker was obtained, $F(4, 16) = 4.14$, $p < .05$. Significantly faster response times were observed for Talker 1 and Talker 5 compared to Talker 3. There was no correlation between the mean latency for a talker and the identification data for that talker that can account for this pattern of response times observed accross talkers. However, there was a systematic effect found in the interaction between week and phonetic environment, $F(8, 32) = 2.44$, $p < .05$. Figure 7 shows the mean response times for each week as a function of phonetic environment. These response times are for correct responses only. It appears that for those environments in which accuracy was high at the outset of training (i.e., for final singletons and clusters), response times became faster each successive week, whereas for those environments in which

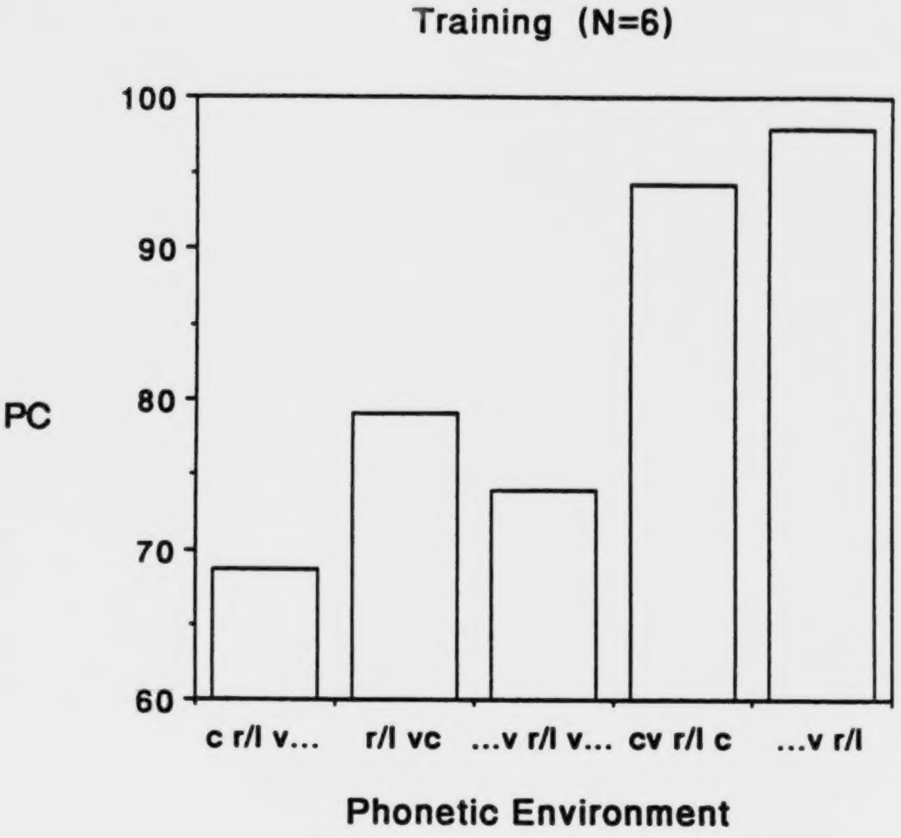


Figure 5. Mean percentage of correct responses in training as a function of phonetic environment.

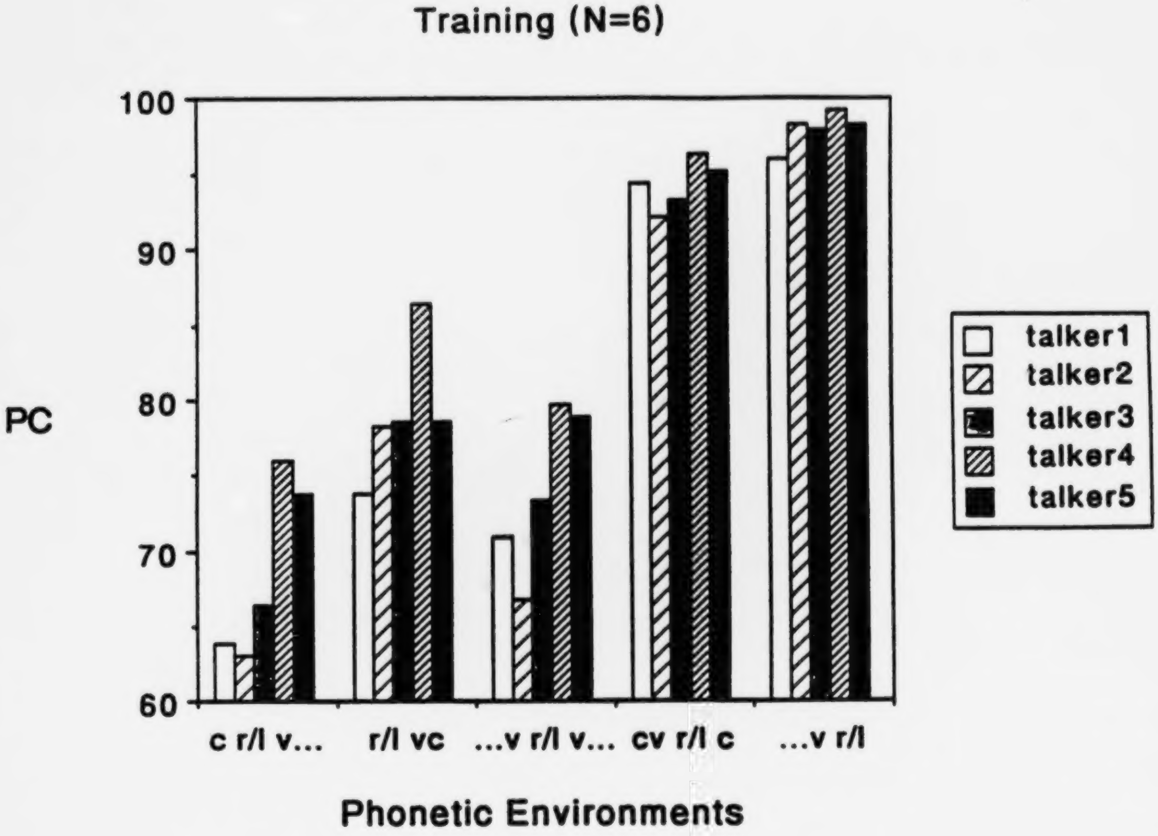


Figure 6. Mean percentage of correct responses in training as a function of talker and phonetic environment.

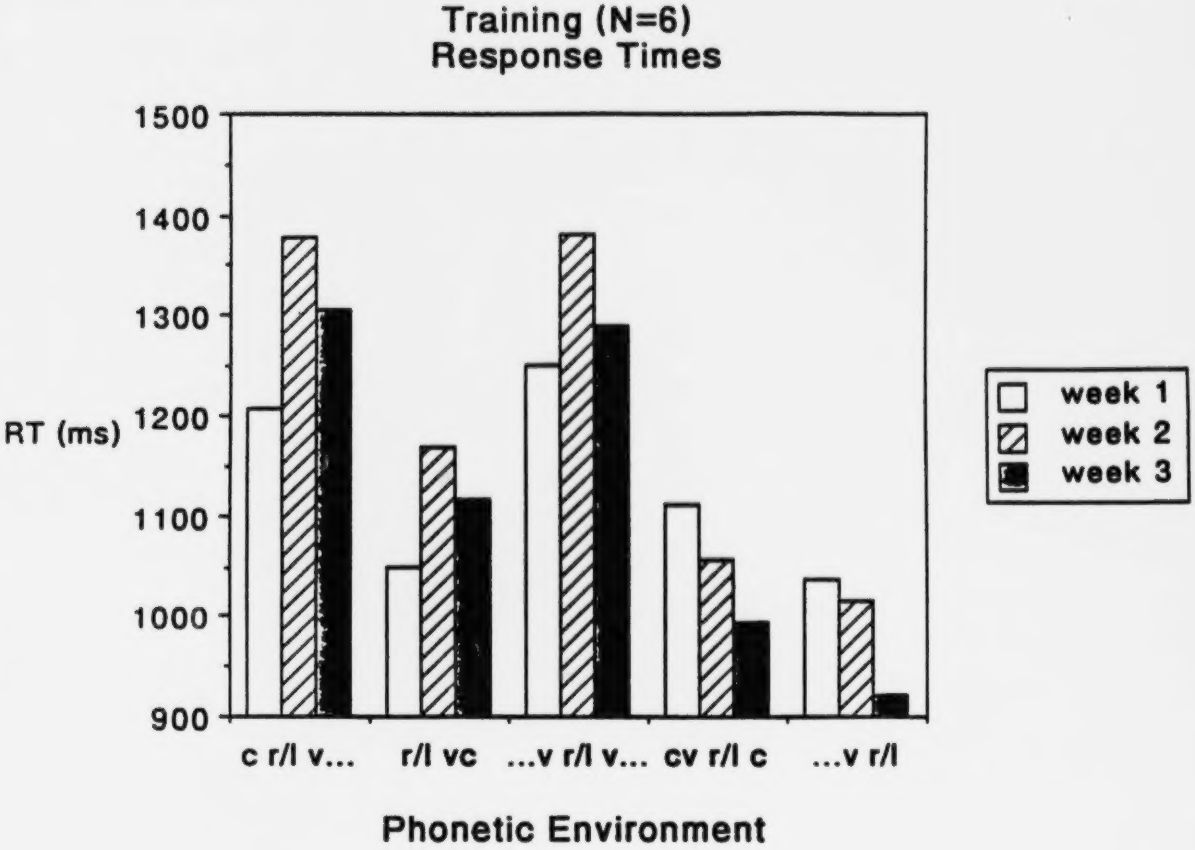


Figure 7. Mean latencies (in ms) for correct responses in training as a function of week and phonetic environment.

accuracy was initially lower, response times became much slower in week 2 than in week 1 and then decreased again in week 3. Thus, the changes in identification performance that occurred from week to week were paralleled by changes in the pattern of response latencies. Moreover, the pattern of response times from week to week varied systematically depending on whether they were from phonetic environments in which identification accuracy was high or from those environments in which accuracy was low.

Insert Figure 8 about here

Generalization

The results of the generalization tests are described next. Figure 8 shows the overall percentage of correct responses for the two generalization tests. Recall that TG1 consisted of novel words produced by a previously unheard talker and TG2 consisted of novel words produced by Talker 4, a talker heard during training. An ANOVA comparing generalization test (TG1-TG2) and phonetic environment for the three subjects who were in both generalization tests yielded no significant effects ($p = .09$) due to insufficient power. However, an examination of Figure 8 suggests that if more subjects had been included in the analysis, performance in TG2 ($M = 83.5$) would have been significantly better than performance in TG1 ($M = 79.9$). This pattern of results implies that subjects were more accurate in their identification of /r/ and /l/ when presented novel words produced by a talker they had heard before than when presented novel words produced by a previously unheard talker.

Discussion

The results of the present experiment indicate that laboratory training procedures can be effectively used to improve the performance of Japanese listeners in identifying /r/ and /l/. Furthermore, generalization to novel stimuli was found to depend, in part, on the relationship between the talkers used during training and those used in generalization. Overall, these results support the claim that for the training of a nonnative phonetic contrast to be robust, the stimuli must contain sufficient variability and the task must bear some resemblance to the actual situations in which the contrast will be used. Earlier work by Strange and Dittman (1984) that used a low uncertainty discrimination task and a small number of synthetic stimuli during training apparently prevented subjects from forming 'robust' phonetic categories that could accomodate the variation in phonetic environments and talkers encountered during generalization, let alone outside the laboratory.

Several of the findings obtained in the present experiment bear further examination. First, the differences in identification performance for /r/ and /l/ in different phonetic environments have now been demonstrated by at least three studies in addition to our own.

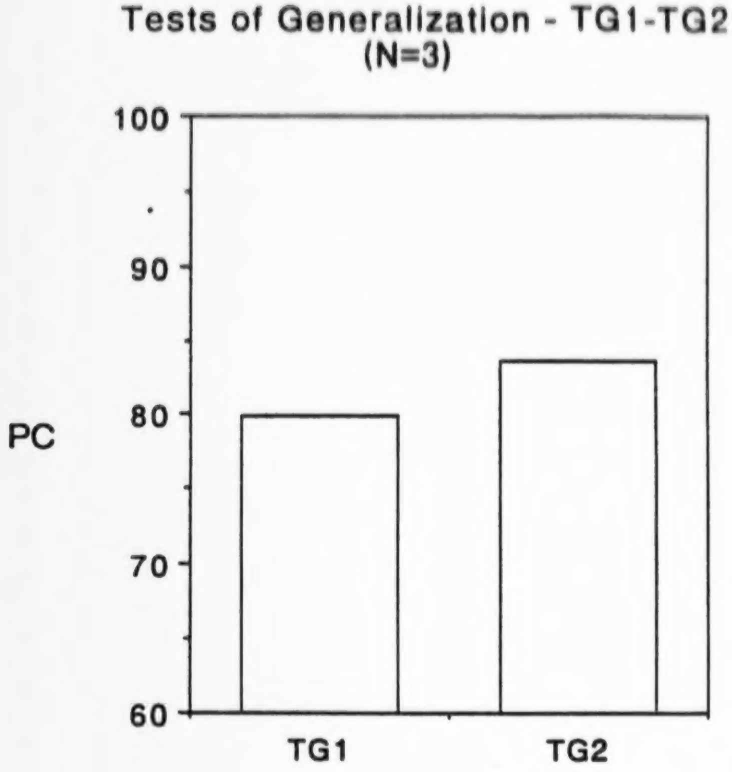


Figure 8. Mean percentage of correct responses in the two tests of generalization, TG1 and TG2.

Yet, no definitive explanation for this finding has emerged. Sheldon and Strange (1982) have proposed that the temporal and spectral characteristics of /r/ and /l/ located in initial environments, especially in consonant clusters, may differ from /r/ and /l/ in final positions. Specifically, they suggest that when /r/ and /l/ "are coarticulated with stop consonants in prevocalic clusters [where performance is worst], their steady-state loci are often not reached or maintained." In final position, however, the acoustic characteristics of /r/ and /l/ tend to influence the formant structure of the preceding vowel, providing additional information about the identity of the liquid in final position. Thus, in the context of initial clusters, acoustic information differentiating /r/ and /l/ may be reduced, whereas in final position, acoustic information differentiating /r/ and /l/ is actually enhanced. According to Sheldon and Strange, these two contexts form the endpoints of a continuum and the "availability and duration" of acoustic features cuing /r/ and /l/ in other phonetic contexts lie between these two extremes.

Partial support for Sheldon and Strange's hypothesis was obtained in an earlier study carried out in our laboratory by Dissosway-Huff, Port, and Pisoni (1982). They found that the duration of /r/ and /l/ in final position was longer than in other phonetic environments. Based on the work of Sheldon and Strange (1982) and Dissosway-Huff et al. (1982), Henly and Sheldon (1986) reasoned that if the effect of phonetic environment on identification of /r/ and /l/ was due solely to acoustic factors, Cantonese listeners, who also have difficulty differentiating /r/ and /l/, should show a pattern similar to the Japanese listeners. Instead, they found a pattern different from that obtained with Japanese subjects. The Cantonese listeners had difficulty with /r/ and /l/ in final position and in initial consonant clusters, and were best at identifying /r/ and /l/ in initial and medial positions. Henly and Sheldon also measured the duration of /r/ and /l/ in their stimuli and found essentially the same durational pattern as obtained by Dissosway-Huff et al. Thus, the relationship proposed by Sheldon and Strange (1982) between the acoustic characteristics of /r/ and /l/ in different phonetic environments and their perception by nonnative listeners was not supported.

Henly and Sheldon suggested that differences between the phonological systems of the two groups were responsible for why the Japanese and Cantonese subjects differed in the pattern of their identification performance. Cantonese listeners were more affected than Japanese listeners because in their phonological system, a variant of /l/, similar to an English /l/, is used in non-final positions. This Cantonese /l/ is used as a "template" to which all other /l/-like sounds are compared. Thus, non-final /r/ and /l/ (except in clusters) can be identified on the basis of either being '/l/-like', in which case the subject identifies the word as containing an /l/, or 'not-/l/-like', in which case the word is identified as containing an /r/. Unfortunately for the Cantonese listener, since English /r/ and /l/ in final position both differ from the Cantonese /l/, the 'template' strategy fails in this environment. However, similarities in the identification performance of Japanese and Cantonese listeners for /r/ and /l/ in initial consonant clusters suggests that, in addition to phonological factors, there are acoustic-phonetic variables at work as well. In short, while the identification of /r/ and /l/ in different phonetic environments by Japanese listeners seems to be closely tied to the acoustic-phonetic characteristics of stimuli, listeners from other language groups, such as Cantonese, may also be influenced by the phonology of their native language.

In order to more fully understand the effect of phonetic environment on the perception of /r/ and /l/, we plan to do an acoustic analysis of the stimuli used in the present experiment. In addition, we plan to carry out perceptual experiments with native speakers of English to see if they show the effects of phonetic environment. At this point, it is not clear whether such effects would be obvious unless the stimuli were degraded since the identification of English words by native speakers of English will undoubtedly be subject to ceiling effects. More sensitive measures, such as response times, may also prove useful to determine if native speakers of English are sensitive to differences among phonetic environments when required to identify words containing /r/ and /l/.

In the present experiment, we also found reliable effects associated with the use of different talkers. Using multiple talkers was originally conceived of as a way to increase the stimulus variability faced by subjects during the training phase of the experiment. We did not anticipate that subjects would be so sensitive to the different talkers used during training, nor did we anticipate the extent to which generalization performance would depend on the relationship between the talkers used during training and the talkers used during generaliztion testing. The work of Mullennix et al. (in press) indicated that, compared to tasks where only one talker was used, the performance of listeners in tasks where the voice of the talker varied from trial to trial was much lower, suggesting the operation of some kind of process in which talker variability is normalized in order for the physical stimulus to be mapped on to an abstract phonetic representation. Mullennix et al.'s results suggested that this process operated at some cost to the cognitive system. However, under most conditions, the costs associated with talker normalization do not seem to noticeably affect perception, at least when the listeners are native speakers of English. In contrast, the results of the present experiment indicate that Japanese listeners presented English words are much more affected by variation in talker than native speakers of English presented with the same stimuli.

The reason why talker variability so dramatically affects the perception of /r/ and /l/ by Japanese listeners is unclear at this time. Despite the fact that the stimuli were pretested with native speakers of English to make sure that all the stimuli were of approximately the same intelligibility, Japanese listeners reliably found some talkers much more intelligible than others. Obviously, some acoustic-phonetic properties underlie the differences between 'good' and 'poor' talkers but until acoustic analyses are completed, we can only speculate about the factors responsible. Durational differences, imperfectly realized formant structure, and the reduction of certain phonemes are all possible candidates at the present time.

There are several implications of the findings obtained in the present experiment regarding talker variation and its effect on perception. First, if the results of the present

231

experiment are any indication, listeners learning a second language seem to be much more sensitive to talker variability than native listeners. Presumably, this is due in part to the ill-formed phonetic categories possessed by these listeners. Second, in order to overcome the effect of talker variability, a large set of talkers may have to be used during training. Apparently, the five talkers used in the present experiment were insufficient to develop robust representations of /r/ and /l/ that the listeners could use in a novel setting. Alternatively, since the listeners did not reach asymptotic performance with the five talkers used during training, it simply may be the case that training subjects until their performance is close to 100% accuracy would result in effective transfer to novel stimuli. Third, it may be the case that training nonnative listeners with some talkers may be more effective in producing new, robust phonetic categories than training with other talkers. That is, subjects may be able to learn to identify nonnative phonetic categories more quickly and accurately using stimuli produced by some talkers than with stimuli from other talkers. Finally, the present results suggest that characteristics of the talker apparently became an integral part of the subject's phonetic representation of /r/ and /l/. The long-standing view that phonetic categories are abstract, canonical representations of speech sounds, may be incorrect because it fails to take into account all the information encoded by a listener. This may be especially relevant for listeners such as nonnative speakers of a language who are likely to have rough nonnative phonetic categories. For these listeners, the information associated with a phonetic category appears to be influenced by stimulus characteristics not included in traditional conceptions of what a segmental representation should be.

In conclusion, the results of the present experiment demonstrate that Japanese listeners can learn to identify /r/ and /l/ in a relatively short time period using a simple laboratory training task that required identification of an item from a minimal pair. However, performance was found to be dependent on the phonetic environment in which the contrast was located and the talkers used during training.The results of two generalization tests showed that listeners apparently learned characteristics of /r/ and /l/ that were not only conditioned by their phonetic environment but were also specific to the talker who produced the items during training. The present experiment raises many interesting and potentially important questions about the nature of stimulus variability in perceptual learning and its role in training nonnative listeners to perceive phonetic contrasts that are not in their language. Further work is currently in progress to address these issues.

References

Aslin, R., & Pisoni, D. (1980). Some developmental processes in speech perception. In G. Yeni-Komshian, J. Kavanagh, & C. Ferguson (Eds.), *Child phonology: Perception and production* (pp. 67–96). New York, NY: Academic Press.

Carney, A., Widin, G., & Viemeister, N. (1977). Noncategorical perception of stop consonants differing in VOT. *Journal of the Acoustical Society of America*, **62**, 961–970.

Dissosway-Huff, P., Port, R., & Pisoni, D. (1982). Context effects in the perception of /r/ and /l/ by Japanese. *Research on speech perception progress report No. 8*. Bloomington, IN: Indiana University.

Gillette, S. (1980). Contextual variation in the perception of L and R by Japanese and Korean speakers. *Minnesota Papers in Linguistics and the Philosophy of Language*, **6**, 59–72.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds "L" and "R". *Neuropsychologia*, **9**, 317–323.

Henly, E. & Sheldon, A. (1986). Duration and context effects on the perception of English /r/ and /l/: A comparison of Cantonese and Japanese speakers. *Language Learning*, **36**, 505–521.

Liberman, A., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, **54**, 358–368.

MacKain, K., Best, C., & Strange, W. (1981). Categorical perception of English /r/ and /l/ by Japanese bilinguals. *Applied Psycholinguistics*, **2**, 369–390.

Miyawaki, K., Strange, W., Verbrugge, R., Liberman, A., Jenkins, J., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of /r/ and /l/ by native speakers of Japanese and English. *Perception and Psychophysics*, **18**, 331–340.

Mochizuki, M. (1981). The identification of /r/ and /l/ in natural and synthesized speech. *Journal of Phonetics*, **9**, 283–303.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (in press). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America.*

Pisoni, D. (1973). Auditory and phonetic codes in the discrimination of consonants and vowels. *Perception and Psychophysics*, **13**, 253–260.

Pisoni, D., Aslin, R., Perey, A., & Hennessy, B. (1982). Some effects of laboratory training on identification and discrimination of voicing contrasts in stop consonants. *Journal of Experimental Psychology: Human Perception and Performance*, **8**, 297–314.

Pisoni, D. & Lazarus, J. (1974). Categorical and noncategorical modes of speech perception along the voicing continuum. *Journal of the Acoustical Society of America*, **55**, 328–333.

Pisoni, D. & Luce, P. (1987). Acoustic-phonetic representations in word recognition. *Cognition*, **25**, 21–52.

Posner, M. & Keele, S. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353–363.

Sheldon, A. & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*, **3**, 243–261.

Strange, W. & Dittman, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception and Psychophysics*, **36**, 131–145.

Webster's Seventh Collegiate Dictionary. (1967). Los Angeles: Library Reproduction Service.

Werker, J. (in press). Becoming a native listener. *American Scientist.*

Werker, J. & Tees, R. (1984). Phonemic and phonetic factors in adult cross-language speech perception. *Journal of the Acoustical Society of America*, **75**, 1866–1878.

II. SHORT REPORTS AND WORK-IN-PROGRESS

F0 Normalization and Adjusting to Talker [1]

Keith Johnson

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported by NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN.

Abstract

Results of two experiments are reported. The experiments involve the presentation of a vowel continuum ("hood"-"hud") at two levels of fundamental frequency (F0) (120 and 240 Hz) and in two types of presentation (items blocked by F0 level and items with different F0 levels randomly intermixed with each other). These presentation conditions correspond to the single-talker/multiple-talker conditions used by Mullennix et al. (in press) and are here called the single-F0 and the mixed-F0 conditions. The experiments test two hypotheses. The first hypothesis is that F0 normalization involves an adjustment-to-talker process in which F0 serves as a cue for speaker identity. This view holds that vowel normalization results from an adjustment of an internal vowel space against which the incoming vowels are compared. The adjustment-to-talker hypothesis is contrasted with a view in which F0 normalization results from the auditory integration of F0 and the harmonics in the F1 region of the vowel spectrum. It is argued that the experiments support the adjustment-to-talker view of F0 normalization. In Experiment 1 there was a large shift in vowel identification (as a function of F0) when items were presented in the mixed-F0 condition, while there was virtually no effect of F0 upon vowel identification when the items were presented in the single-F0 condition. This contrast effect in F0 normalization has not been reported before. In Experiment 2, the contrast effect was found again. In addition, reaction time data was comparable to that found by Mullennix et al. (in press) for the single-talker vs. multiple-talker presentation of naturally produced words. Vowel identification in the single-F0 condition was on average 50 ms faster than it was in the mixed-F0 condition. Further analysis revealed that the reaction time disadvantage in the mixed-F0 condition occurred for those items which involved a change in F0 from one token to the next. These data indicate that hearers make a perceptual adjustment upon encountering a new voice. The auditory theory of F0 normalization does not predict these results. They are, however, predicted by a vowel-space-adjustment model of vowel normalization.

F0 Normalization and Adjusting to Talker*

It is well documented that hearers are influenced by fundamental frequency (F0) when they make judgements about vowel quality (Miller, 1953; Fujisaki & Kawashima, 1968; Slawson, 1968; Ainsworth, 1975; Traunmüller, 1981). Traunmüller (1981) has suggested that F0 normalization is the result of the auditory integration of F0 and the harmonics in the region of the first oral resonance (F1) which results in a shift of the perceived F1 (F1'). Sussman (1986) proposes that perceptual vowel normalization is achieved by a specialized component of the auditory system which is sensitive to the ratio of F1 to F0. He notes that, in the auditory systems of some species of bat, there are nerve cells which are sensitive to frequency ratios. Syrdal and Gopal (1986) also include reference to the relationship of F1 to F0 in their vowel classification scheme. Instead of looking ant the ratio of F1 and F0, however, they define the relationship in terms of difference between Z1 and Z0 (the Bark transformed values of F1 and F0). If Z1 minus Z0 is greater than 3 Bark the vowel is classified as open and if the difference of Z1 and Z0 is less than 3 Bark the vowel is classified as closed. They note that the use of Bark-difference definitions of vowel features accomplishes speaker normalization. All three versions of an auditory account of F0 normalization avoid any reference to the role of F0 in speaker-identity (or of the role of speaker-identity in vowel normalization) and so give F0 normalization an ontological status different from other processes of speaker normalization such as contextual vowel normalization (Ladefoged & Broadbent, 1957).

I have argued elsewhere (Johnson, 1988a,1988b) that this view of F0 normalization is incorrect, and that a part of the influence which F0 has on vowel identification stems from the fact that F0 is a cue for speaker-identity. As a cue for speaker-identity, F0, in this view, is used in a range normalization process (Gerstman, 1968) to estimate speaker-dependent formant ranges. I will call this the adjustment-to-talker view. The model of vowel normalization described by Bladon, Henton, and Pickering (1984) is one example of the adjustment-to-talker approach. This is not at first obvious because they describe the model as "auditory", but it is actually a two stage model. They propose that vowel normalization is accomplished by shifting the auditory spectra of vowels produced by female speakers down by 1 Bark before comparing them to spectral templates based on vowels produced by men. The auditory stage in the model is in the calculation of "auditory" spectra (Bladon & Lindblom, 1981). The second stage (when spectra are shifted along the Bark scale) involves an adjustment-to-talker.

The experiments reported here provide a test of these hypotheses. Before discussing the present experiments, I will review some recent research on the perceptual consequences of having to adjust to different talkers during word recognition.

In some recent research concerning word recognition under conditions of talker-variability, Mullennix, Pisoni, and Martin (in press) found that when hearers were required to identify words in different levels of noise, word recognition performance was impared by random variation of talker identity (41% correct in the single-talker condition vs. 34% correct in the multiple-talker condition). In the single-talker condition all of the words presented for identification were produced by the same talker; in the multiple-talker condition the identity of the talker (out of a pool of 15 talkers) varied ramdomly from word to word (hence the multiple-talker condition can be called a talker-variability condition). They found a similar effect of talker-variability with tokens which had been degraded by the introduction of signal correlated noise (Horri, House, & Hughes, 1971). Identification performance was 69% correct vs. 48% correct in the single-talker and mixed-talker conditions, respectively. Mullennix et al. also found reliable reaction time differences between single-talker and mixed-talker conditions in two naming experiments. Subjects could repeat aloud auditorily presented words in the single-talker condition about 50 ms faster than they could in the mixed-talker condition (averaged over lexical density and word frequency conditions in two experiments). These data were interpreted as indicating that hearers must adjust to talker in the multiple-talker condition while this adustment is not required in the single-talker case.

Previous research in F0 normalization has involved the presentation of synthetic speech tokens in what is essentially a multiple-talker condition, though the "multiple-talker" condition has usually been composed of only two levels of F0 (Miller, 1953 ; Fujisaki & Kawashima, 1968; Slawson, 1968; and Traunmüller, 1981). In the experiments reported here synthetic speech tokens from a "hood"-"hud" continuum were presented at two levels of F0 and in both single-talker and multiple-talker conditions.

Experiment 1

Subjects identified the vowels from two versions of a "hood"-"hud" continuum in two types of presentation. There were two versions of the vowel continuum, one with high F0 (240 Hz) and one with low F0 (120 Hz). In the single-F0 condition the tokens were blocked by F0, thus the F0 of the tokens was entirely predictable within blocks. In the mixed-F0 condition the tokens of the two F0 continua were randomly intermixed with each other. The single-F0 condition corresponds to the single-talker condition of Mullennix et al. (in press) and the mixed-F0 condition is analogous to their multiple-talker condition, albeit with only two talkers.

Method

Subjects. Twenty undergraduate students at Indiana University participated in the experiment (12 female, 8 male). All were native speakers of American English who had never experienced any speech or hearing deficiences. They received partial course credit in an introductory psychology course for their participation.

Materials. The stimuli used in this experiment were synthetic CVC syllables in a vowel continuum from [hʊd] to [hʌd]. Two continua were synthesized (using the Klatt, 1980 cascade-parallel formant synthesizer) - one with a steady-state F0 of 120 Hz, the other with steady-state F0 of 240 Hz. The formant values used in synthesizing the vowels are shown in Table 1. These formant values had been used in a previous study of vowel F0 normalization (Johnson, 1988a). The syllables were 285 ms in duration. The aspiration noise of the /h/ was 95 ms long (with F1 and F2 slightly higher than the F1 and F2 of the vowel as naturally occurs as a result of tracheal coupling). The steady-state vowel portion of the stimuli was 160 ms long. Bandwidths of F1-F3 were 110, 75, and 110 Hz, respectively. F4 and F5 were 3500 Hz and 4200 Hz, both with a bandwidth of 300 Hz, and were steady-state throughout the syllables. The final transitions into /d/ were 30 ms long and ended at 300,1700 and 2516 Hz for F1-F3, respectively. The F3 transition of all tokens dipped to 2116 over the first 15 ms and then rose to 2516. The low F0 continuum was synthesized with a steady-state F0 of 120 Hz, while the high F0 continuum had a steady-state F0 of 240 Hz. The peak amplitude relations among tokens reflected the intrinsic amplitudes found in natural speech (Lehiste and Peterson, 1959). Peak amplitude (expressed in dB relative to the maximum value possible in the synthesizer) of the low F0 continuum ranged from -9.1 dB for [ɷ] to -7.1 for [ʌ], while peak amplitude for the high F0 items ranged from -2.6 [ɷ] to -0.7 dB [ʌ]. The tokens at the "hud" end of the continuum were slightly louder than the tokens at the "hood" end and the high F0 tokens were louder (by 6.45 dB on average) than the low F0 tokens.

Insert Table 1 about here

Procedure. The two vowel continua were presented in random order to each group of subjects (at a listening level of 80 dB for the [ʌ] token in the high F0 continuum) in two types of presentation. In the single-F0 condition the tokens in each continuum were blocked by F0. In the mixed-F0 condition the tokens from the two continua were randomly intermixed with each other. Subjects were randomly divided into two groups. One group of subjects responded first to the items in the single-F0 condition and then to the same items again in the mixed-F0 condition. The other group heard the items in the mixed-F0 presentation type first and then responded to them again in the single-F0 condition. The groups will be called the mixed-first and single-first groups. In the single-F0 condition, the order of presentation of F0 level was counter-balanced across subjects. So, presentation type and F0 level were treated as within subjects variables while order of presentation was treated as a between subjects variable.

A CRT was mounted at approximately eye level for each subject. The words "HOOD" and "HUD" were presented on the monitor at the left and right bottom corners of the screen. Subjects were told to use these labels as an indication of which button to press in a forced choice identification task. One half of each subject's responses in each condition were collected with "HOOD" as the right-hand response and "HUD" as the left-hand response, and

Table 1

Formant values of the test tokens used in Experiments 1 and 2.

Token #	1	2	3	4	5	6	7
F1	474	491	509	526	543	561	578
F2	1111	1124	1137	1150	1163	1176	1189
F3	2416	2424	2432	2440	2448	2456	2464

for the other half of the trials the right-hand response was "HUD" and the left-hand response was "HOOD". Button to response associations were switched at intervals of 70 trials, so within a block of 70 trials the association was constant.

Each token in the two continua was presented 10 times in each of the two types of presentation. Thus, the number of observations per subject was 280 (7 tokens * 2 F0 levels * 2 presentation types * 10 presentations). The experiment was conducted online by a PDP 11/34 mini-computer at the Speech Research Laboratory at Indiana University. Subjects were run in groups of five.

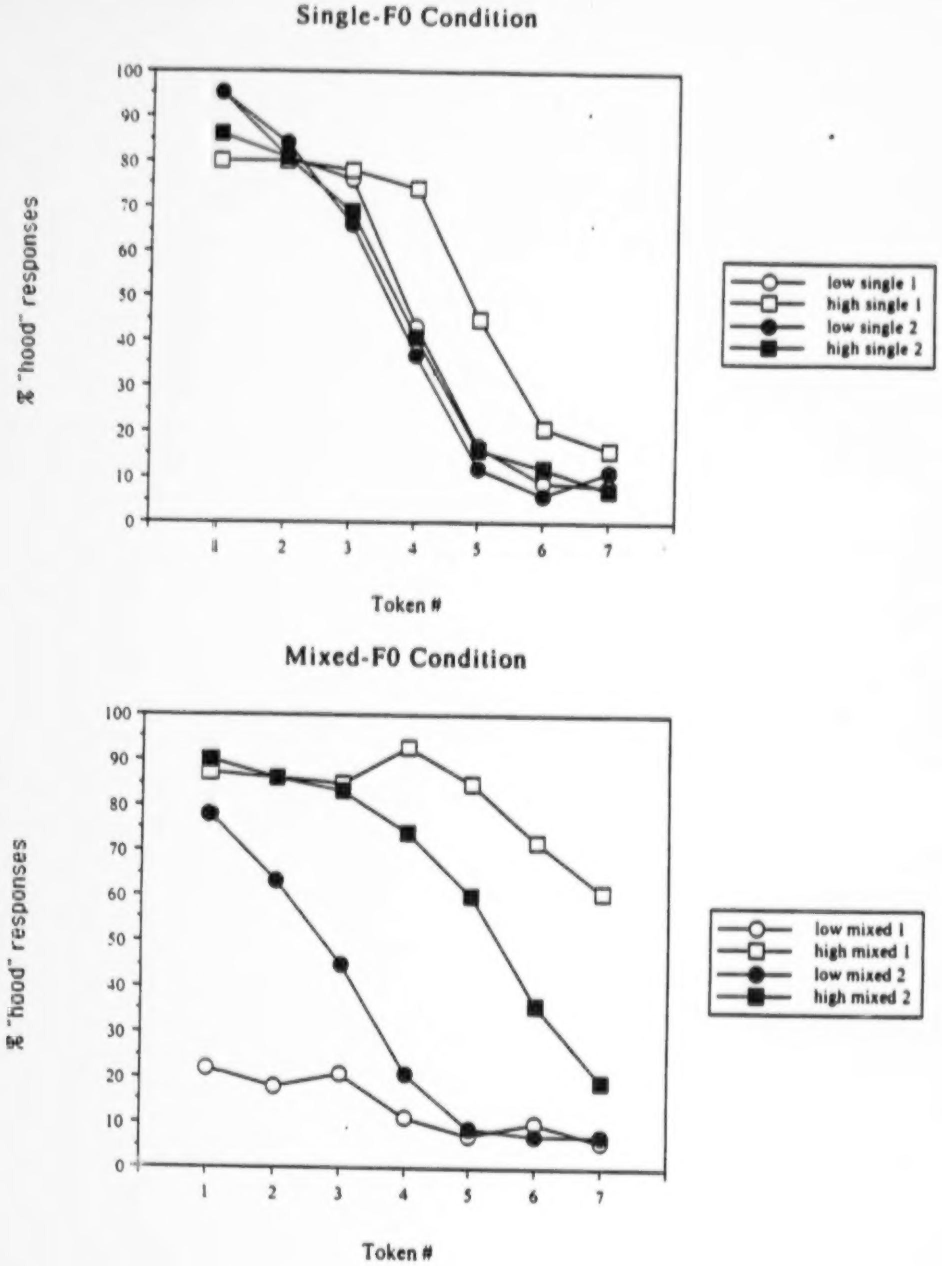
Results

The results of Experiment 1 are shown in Figure 1. The single-F0 and mixed-F0 conditions are in the top and bottom graphs, respectively. In each graph the identification functions for subjects in the single-first group are plotted with solid symbols and the responses of the subjects in the mixed-first group are plotted with open symbols. Responses to the high F0 continuum are plotted with squares and responses to the low F0 continuum are plotted with circles. These data were analyzed in a four-way repeated measures ANOVA with factors PRESENTATION TYPE (mixed-F0 vs. single-F0), F0 LEVEL (120 Hz vs. 240 Hz), GROUP (mixed-first vs. single-first) and TOKEN (1-7).

Insert Figure 1 about here

There were two main effects; TOKEN and F0 LEVEL ($[F(6,108) = 118.06, p < .0001]$ and $[F(1,18) = 76.7, p < .0001]$, respectively. Both of these effects are unsurprising. Of course, we expect vowel identification to vary as a function of token number, also, previous research on F0 normalization would lead to the prediction that vowel identification would vary as a function of F0 level. Low F0 items were labeled "hood" 34.5% of the time while 61.5 % of the high F0 items were identified as "hood".

The interaction which is of primary interest here is the one between PRESENTATION TYPE and F0 LEVEL. This interaction, which was highly significant $[F(1,18) = 65.79$, $p < .0001]$, is illustrated in Table 2. The data in this table are presented as percent "hood" responses averaged across the tokens in the continua and averaged across subjects. As indicated in Table 2 and Figure 1 subjects showed a very large effect of F0 level on vowel identification responses if the items were presented in the mixed-F0 condition, and very little effect of F0 level when the items were presented blocked by F0.



Single-F0 Condition

Mixed-F0 Condition

Figure 1. Identification data from Experiment 1 as a function of token number, presentation type, subject group and F0 level. The mixed-first group is plotted with open symbols. The single-first group is plotted with closed symbols. The high F0 continuum is plotted with squares and the low F0 continuum is plotted with circles.

Insert Table 2 about here

There were several other significant interactions which have less of a bearing upon the hypotheses being tested. First, there was an interaction between PRESENTATION TYPE and TOKEN [$F(6,108) = 30.21$, $p < .0001$]. This interaction is a result of the fact that the F0 normalization effect was present in the mixed-F0 condition and not in the single-F0 condition. For example, if we average the four functions in the top graph of Figure 1 the percent "hood" responses for token #1 is 89%, but the average for token #1 in the mixed-F0 condition is only 69%. The effect of low F0 on the perceptual identity of token #1 in the mixed condition was to reduce the number times this token was identified as "hood". Similarly, the effect of high F0 on the perceptual identity of token #7 (again only in the mixed-F0 condition) was an increase in the number of "hood" responses to this token. Thus, the identification functions (averaged over F0 level and subject groups) are flatter in the mixed-F0 condition than they are in the single-F0 condition.

Also, in Figure 1 it is obvious that the effect of F0 level in the mixed-F0 condition was not as large for the subjects in the single-first group. This is borne out in the ANOVA as an interaction between GROUP and F0 LEVEL [$F(1,18) = 13.67$, $p < .01$] and as a three way interaction of GROUP, PRESENTATION TYPE and TOKEN [$F(6,108) = 7.82$, $p < .0001$]. For those subjects who heard the single-F0 condition first the difference in vowel quality as a function of F0 was less extreme than it was for the subjects who first responded to the items in the mixed-F0 condition. The three way interaction indicates that when responses are averaged over F0 the performance of the mixed-first group on mixed-F0 trials is flatter than block first performance on mixed-F0 trials. What seems to be happening here is that, when the items are presented in the single-F0 condition before being presented in the mixed-F0 condition, subjects are able to hear out exemplars of "hood" and "hud" in both of the continua. Then, when the items are presented in the mixed-F0 condition, the subjects are more likely to identify instances of "hud" in the high F0 continuum and instances of the "hood" in the low F0 continuum. As Figure 1 illustrates, those subjects who responded to the mixed-F0 condition first had a tendency for greater separation of the two F0 continua than did those subjects in the single-first group.

Another statistical effect which results from the difference between groups is the GROUP by TOKEN interaction [$F(6,108) = 7.72$, $p < .0001$]. This effect appears to be a "shadow" of the GROUP by TOKEN by PRESENTATION TYPE interaction which was discussed in the preceding paragraph, although the greater number of "hood" responses to the high F0 items by subjects in the mixed-F0 first group in the single-F0 condition is also a part of this interaction (note that the four way interaction which includes F0 LEVEL as a factor was not significant).

Two other interactions were significant in the ANOVA. There was an interaction between

Table 2

The interaction of PRESENTATION TYPE and F0 LEVEL in Experiment 1. The data in this table are percent "hood" identifications as a function of presentation type and F0 level averaged across subjects and tokens.

	Low F0	High F0
Single-F0	45.7	50.4
Mixed-F0	23.2	72.6

F0 LEVEL and TOKEN NUMBER [$F(6,108) = 11.09$, $p < .0001$]. This interaction indicates that the effect of F0 level on vowel identification was not the same for each of the tokens in the continuum. Tokens in the middle of the continuum were more affected by the manipulation of F0 than were the tokens at the ends of the continuum (i.e. a boundary shift vs. a global change).

And finally there was an interaction between PRESENTATION TYPE and GROUP [$F(1,18) = 10.3$, $p < .01$]. Averaged across all other conditions the mixed-first group identified 47% the tokens in the mixed-F0 condition as "hood", and 52% the tokens in the single-F0 condition as "hood". Subjects in the single-first group identified 48% of the tokens in the mixed-F0 condition as "hood" and 45% of the tokens in the single-F0 condition as "hood". This effect can be summarized by saying that there were fewer "hood" responses in the first presentation condition for each group. This interaction has no bearing on the hypotheses under investigation.

Discussion

Both the auditory view (Traunmüller, 1981; and Sussman, 1986) and the adjustment-to-talker view predict that an F0 normalization effect will be exaggerated in the mixed-F0 condition as opposed to the single-F0 condition. In the auditory view of F0 normalization, an auditory contrast effect (Crowder, 1981; Fox, 1985) may increase the spectral differences between tokens with different F0 levels in the mixed-F0 condition and so increase the vowel quality differences between the two synthetic continua. This account of contrast has the following features: (1) vowel representations differ as a result of the auditory processing of vowels (and for no other reason), (2) vowel contrast in the mixed-F0 condition results from auditory contrast (e.g. Crowder's (1981) model of lateral inhibition across vowel spectra).

The adjustment-to-talker model also predicts a contrast effect in the mixed-F0 condition, but at a different level of analysis. Here, instead of contrasting at the level of auditory vowel representations, the contrast effect is at the level of speaker-identities. In this view, the degree of perceived speaker difference (as a function of F0 difference) will be greater in the mixed-F0 condition because of a relatively high-level contrast effect (Parducci, 1965, 1975). Consequently, if the perceived difference between speakers in the mixed-F0 condition is greater in the mixed-F0 condition, the degree of difference in vowel identification will be greater as well.

So, this experiment, although it demonstrates and interesting (and until now, unreported) phenomenon, does not really separate the two hypotheses of vowel normalization. Both hypotheses predict a contrast effect. A second experiment was performed in order to separate the two hypotheses. In this second experiment reaction time data as well as identification data were collected. This experiment separates the hypotheses because the auditory explanation of vowel normalization predicts that there will be no reaction time difference between the single-F0 and mixed-F0 conditions, while the adjustment-to-talker hypothesis does predict a reaction time difference.

Experiment 2

Experiment 2 is identical to Experiment 1 in all respects except that the peak amplitudes of the tokens were equated (to avoid possible effects of amplitude variation on reaction time), and reaction time measurements were collected.

The adjustment-to-talker view of F0 normalization predicts that, when the F0 of tokens is unpredictable within an experimental block, subjects will have to adjust to a new talker each time they encounter an item with F0 different from the F0 of the preceding item. Mullennix et al. (in press) found that when talker voice varies from item to item in a word naming task just such a reaction time effect is observed. There are several differences between the present experiment and that of Mullennix et al. which suggest that the reaction time difference may not be found in this experiment. First, Mullennix et al. used naturally produced monosyllabic words. The differences between talkers in their study therefore included F0, spectral slope, other aspects of voice quality such as laryngealization and breathiness, and differences in formant ranges. The tokens used in the present experiment, on the other hand, differed (from one continuum to the next) only in F0. In a previous experiment, I found that a difference of 50 Hz was enough to cause listeners to classify tokens as having been produced by different talkers (Johnson, 1988b), but clearly the differences between "talkers" in this experiment were much less than those between talkers in the Mullennix et al. study. Second, the multiple-talker condition in the Mullennix et al. study involved the presentation of tokens which had been produced by 15 different talkers. In the present study only two different levels of F0 were presented in the mixed-F0 condition. So, this experiment is a rather severe test of the adjustment-to-talker account of F0 normalization.

The auditory F0 normalization hypothesis predicts that there will be no reaction time difference between the two presentation conditions (single-F0 and mixed-F0) because the model holds that both the F0-normalization effect and the contrast effect which were observed in Experiment 1 are auditory in nature. In other words, the auditory hypothesis holds that exactly the same cognitive processing is taking place in both of the presentation conditions. In both conditions, incoming vowels are being identified by whatever mechanism we care to posit and the differential effects of F0 level and stimulus context are the result of auditory processing. Thus, the auditory hypothesis predicts that there will be no reaction time difference for the two presentation conditions.

Method

Subjects. Twenty-four undergraduate students at Indiana University participated in the experiment (18 female, 6 male). All were native speakers of American English who had never experienced any speech or hearing deficiencies. They received partial course credit in an introductory psychology course for their participation.

Materials and Procedure. Two vowel continua were synthesized using the same formant values, bandwidths and durational qualities as those used in Experiment 1. Peak amplitude was digitally equated across all tokens.

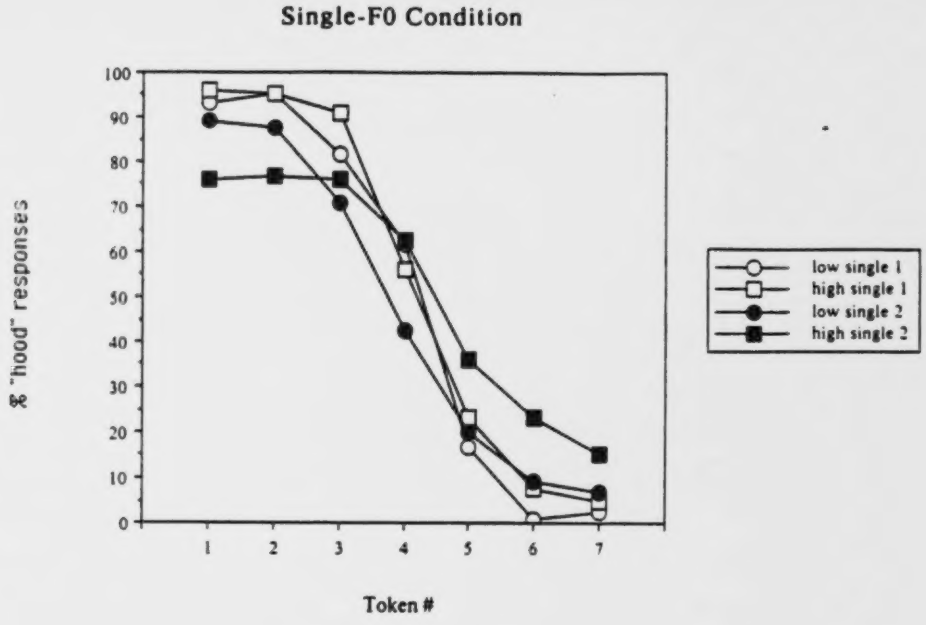
The procedure was identical to that used in Experiment 1. Subjects were tested in groups of six; each group being tested in a different (counter-balanced) ordering of the conditions.

Results

Identification Data. The identification data from Experiment 2 are shown in Figure 2. As in Figure 1 the two presentation conditions are shown in separate graphs, the responses of the mixed-first group are plotted with open symbols in both graphs, while the responses of the single-first group are plotted with filled symbols. Responses to the low F0 continuum are plotted with circles and the responses to the high F0 continuum are plotted with squares. The identification data were analyzed in a four way repeated measures ANOVA with factors PRESENTATION TYPE (mixed-F0 vs. single-F0), F0 LEVEL (120 Hz vs. 240 Hz), GROUP (mixed-first vs. single-first) and TOKEN (1-7).

Insert Figure 2 about here

As in Experiment 1 the only main effects in the analysis were for F0 LEVEL [$F(1,22) = 91.76, p < .0001$] and TOKEN [$F(6,132) = 154.14, p < .0001$]. The low F0 continuum was identified as "hood" 35.5% of the time while 64.9% of the high F0 items were labeled "hood". Also, as in Experiment 1, the interaction of PRESENTATION TYPE and F0 LEVEL was highly significant [$F(1,22) = 123.07, p < .0001$]. The effect of F0 on vowel identification (when tokens with different F0 are presented in blocks vs. randomly intermixed with each other) can be observed in Figure 2 as the difference between the top and bottom graph. The average percent "hood" responses to each continuum (low and high F0) are shown in Table 3. When items which differed in F0 were presented intermixed with each other there was a large F0 normalization effect, while when the items were presented in separate blocks no such effect obtained. The three way interaction of PRESENTATION TYPE, F0 LEVEL and TOKEN was also significant [$F(6,132) = 4.82, p < .001$]. This interaction indicates that the effect of F0 in the mixed-F0 condition was greater for some tokens in the continuum than



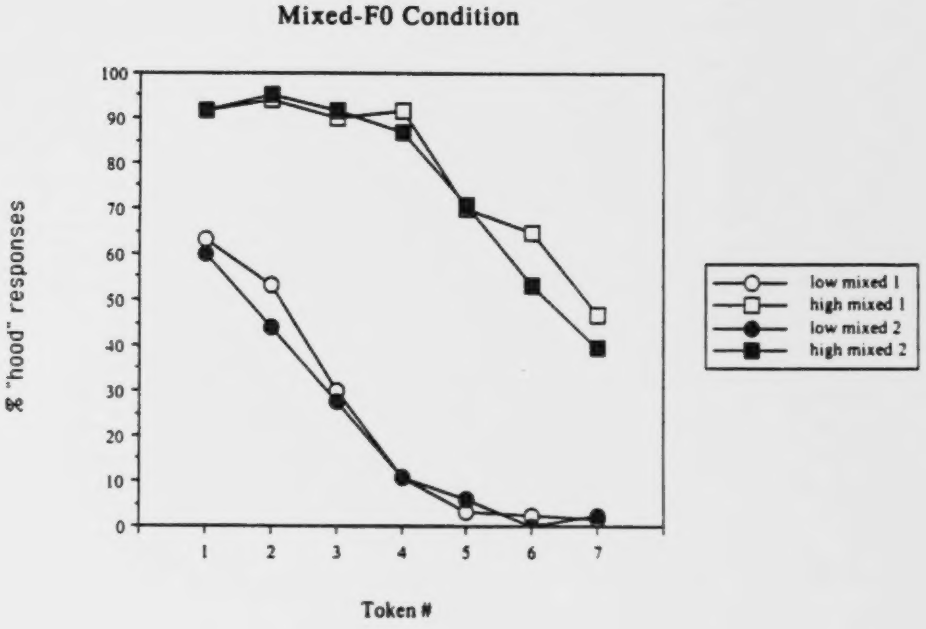


Figure 2. Identification data from Experiment 2 as a function of token number, presentation type, subject group and F0 level. The mixed-first group is plotted with open symbols. The single-first group is plotted with closed symbols. The high F0 continuum is plotted with squares and the low F0 continuum is plotted with circles.

for others. In particular, tokens at the "hud" end of the continuum were affected by high F0 and tokens at the "hood" end of the continuum were affected by low F0.

Insert Table 3 about here

As in Experiment 1, there was an interaction between PRESENTATION TYPE and TOKEN [$F(6,132) = 32.61, p < .0001$] which indicates that the average identification function in the mixed-F0 condition was flatter than was the average identification function in the single-F0 condition. Also, as in Experiment 1, there was an interaction between F0 LEVEL and TOKEN [$F(6,132) = 15.33, p < .0001$], which indicates that the effect of F0 (averaged over groups and presentation conditions) was a boundary shift and not a global change in probability of "hood" response.

Finally, there were two significant interactions which involved group differences. As is clear in Figure 2, the two groups of subjects (mixed-first and single-first) had virtually identical response functions in the mixed-F0 condition. While in the single-F0 condition the response functions for the single-first group were somewhat flatter than those of the mixed-first group. This difference was reflected in the PRESENTATION TYPE by TOKEN by GROUP interaction [$F(6,132) = 4.25, p < .001$]. Also, from the graph of the single-F0 condition functions it appears that this group difference was larger for the high F0 level than for the low F0 level (i.e. the function for the high F0 continuum is flatter than the function for the low F0 continuum). This was reflected in the four way interaction of PRESENTATION TYPE, GROUP, F0 LEVEL and TOKEN [$F(6,132) = 3.02, p < .01$]. It is not clear why the subjects in the single-first group would show less categorical identification of the continua in the single-F0 condition.

Reaction Time Data. Average reaction times were also analyzed in a repeated measures ANOVA with factors F0 LEVEL, PRESENTATION TYPE and GROUP. The values that were entered into this analysis were averaged across both response ("HOOD" or "HUD") and token. The reaction time measurements indicate response latency from item onset. The only statistically significant effect was a the main effect for PRESENTATION TYPE [$F(1,22) = 5.39, p < .05$]. Average reaction time in the mixed-F0 condition was 697 ms and in the single-F0 condition was 647 ms. This 50 ms difference is comparable to the reaction time difference found by Mullennix et al. (in press) for naming latency in multiple-talker versus single-talker conditions. The main effect for F0 LEVEL approached significance [$F(1,22) = 3.76, p = .0653$]. The trend was for items with low F0 to be identified more quickly than the items with high F0. This effect may relate to the relative naturalness of the different levels of F0. The voice source of the synthesizer (as is true with many synthesizers) sounds more natural at lower F0 levels.

Table 3

The interaction of PRESENTATION TYPE and F0 LEVEL in Experiment 2. The data in this table are percent "hood" identifications as a function of presentation type and F0 level averaged across subjects and tokens.

	Low F0	High F0
Single-F0	48.4	52.7
Mixed-F0	22.6	76.9

An additional analysis of the reaction time data from the mixed-F0 condition was performed to assess the effect of changing F0 from token to token. The reaction time data were classified by F0 of the item being identified and by the F0 of the immediately preceding item. This classification scheme results in four classes of reaction times; low F0 items which were immediately preceded by a low F0 item, low F0 items which were immediately preceded by a high F0 item, high F0 items which were immediately preceded by a low F0 item and high F0 items which were immediately preceded by a high F0 item. These data for each of the 24 subjects were entered into a three-way repeated measures ANOVA with factors: TOKEN F0 (high or low), CONTEXT F0 (high or low) and GROUP (mixed-first or single-first). There was a main effect of TOKEN F0 [$F(1,22) = 9.96, p < .01$]. This is consistent with the marginal F0 effect found in the overall analysis. There was also a significant interaction between the TOKEN F0 and CONTEXT F0 factors [$F(1,22) = 5.98, p < .05$]. This interaction is illustrated in Figure 3. In a *post hoc* comparison of means it was found that the effect of context was significant for the high F0 items, but not for low F0 items.

Insert Figure 3 about here

Discussion

As in Experiment 1, the pattern of identification responses found in this experiment clearly indicate that presentation type determines whether the shift in identification associated with F0 normalization will occur. When vowel tokens are presented in what amounts to a mixed-talker situation subjects tend to identify vowels with high F0 more as "hood" and vowels with low F0 more as "hud". This is the expected effect of F0 on vowel identification given some process of F0 normalization, and this pattern of results has been reported widely (Miller, 1953; Fujisaki & Kawashima, 1968; Slawson, 1968). What is unique and intriguing about the present finding is that this pattern of identification for continua with different F0 is not present when the contunua are presented in separate blocks.

The difference in reaction time between the mixed-F0 and single-F0 conditions which was observed in this experiment (50 ms) is comparable to the reaction time difference observed by Mullennix et al. (in press) between their single-talker and multiple-talker conditions. This similarity across experiments suggests that the same mechanism is involved. One possible conclusion is that the adjustment to talker to which Mullennix et al. attributed the reaction time difference includes the F0 normalization effect which was found in the mixed-F0 condition in this experiment.

General Discussion

One striking and totally unexpected result of Experiment 2 was the lack of group differences. Comparing Figures 1 and 2 it is obvious that the differences between subject groups

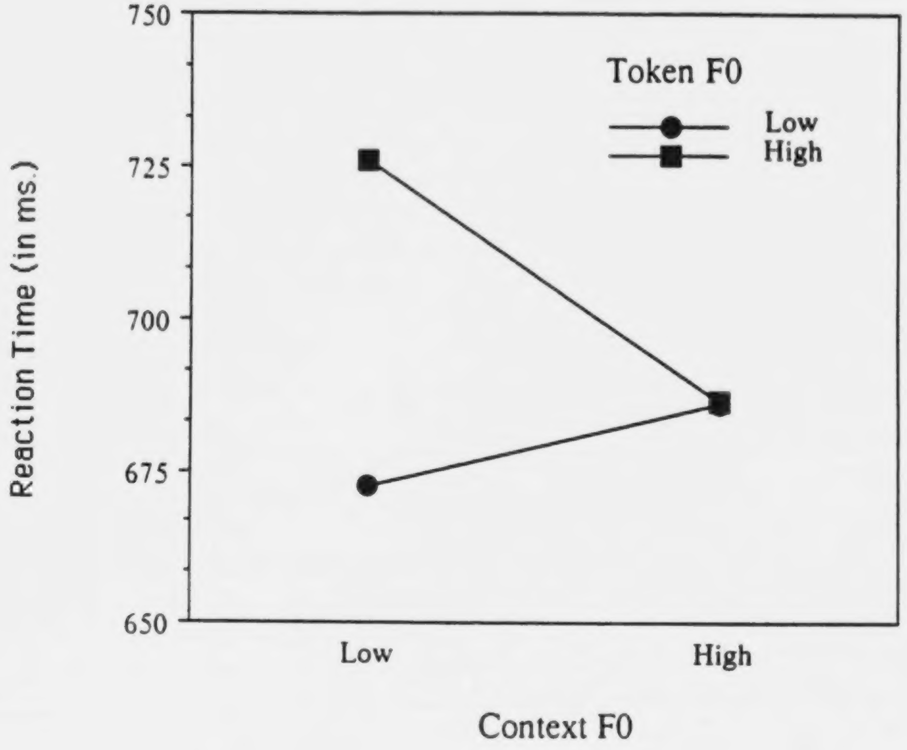


Figure 3. Reaction time data in the mixed-F0 condition in Experiment 2 as a function of F0 level of the token and F0 level of the immediately preceding token

which were present in Experiment 1 have all but vanished in Experiment 2. This is a striking difference because the only difference between the two experiments was that in Experiment 2 the amplitudes of the tokens were digitally equated in order to get valid reaction time measurements (without confounding amplitude and F0). Any effect of this manipulation on identification behavior was totally unexpected.

It may be that the amplitude differences between continua in Experiment 1 served to magnify the F0 difference between continua for the mixed-first group, while the single-first group had an opportunity (in the single-F0 condition) to "hear out" examplars of both vowel categories in the two continua, and so were less affected by the amplitude difference in the mixed-F0 condition. This is the interpretation given in the discussion of the results of Experiment 1. Evidently, the removal of this confounded physical difference (in Experiment 2) was enough to allow subjects in the mixed-first condition to also "hear out" instances of both vowel categories in the mixed-F0 condition.

It is not clear if this pattern of results bears on the auditory or adjustment-to-talker views of F0 normalization. One possible interpretation, from the adjustment-to-talker view, is that, by confounding amplitude with F0, the degree of perceived talker difference was increased for the mixed-first group in Experiment 1. The fact that this coupling of cues did not affect the single-first group as well may be the result of being able to ignore amplitude differences given prior exposure to the two voices. This interpretation, although a possibility, is not the primary focus of the present experiments.

The primary results of these two experiments can be summarized as: (1) A large and stable contrast effect in F0 normalization has been found. When tokens from vowel continua with different F0 are presented randomly intermixed with each other there is an effect of F0 upon vowel identification, however, when tokens are presented blocked by F0 the effect of F0 is severely diminished. This contrast effect in F0 normalization has not been reported before. It was argued that both an auditory view of F0 normalization and an adjustment-to-talker view of F0 normalization could account for this finding. (2) Reaction time data have been reported which suggest that hearers must make an adjustment-to-talker for items which differ in F0 from their immediate context. This finding is predicted by the adjustment-to-talker view of F0 normalization and not by the auditory view of F0 normalization.

Finally, let's consider how these data relate to Bladon et al.'s (1984) model of vowel normalization. In their model, vowel normalization involves the adjustment of the spectra of vowels produced by females in oder to make them comparable to vowel category templates which are based on the spectra of vowels produced by males.[1] This model predicts that an adjustment will have to be made for female voices and not for male voices, or in the more general case (if the vowel category templates are midway between spectra for men and women), the model predicts that an adjustment will have to be made for all vowel tokens. The reaction time data of Experiment 2 are not consistent with this model. These data (Figure 3) indicate that an adjustment-to-talker is only made when there is a change in voice. The spectrum-sliding model of Bladon et al. entails that the spectral templates are fixed and that the representation of the incoming vowel must be adjusted to make a comparison with the template. The vowel space adjustment model which I am proposing here holds that, upon detecting a change in speaker (or at the onset of an exchange), the hearer adjusts an internal vowel space and that the representations of incoming vowels are then evaluated relative to this vowel space without being altered in preparation for the comparison.

---

[1] There is no reason why the vowel category templates couldn't be based on spectra of female vowels, or on spectra close to the average of male and female vowel spectra.

This is not to say that listeners evaluate vowels by their formant frequencies (in Hz). Of course, the representations of vowels are auditory representations. I have reported elsewhere (Johnson, 1988b) data which suggest that some portion of the shift in identification which is found in F0 normalization experiments (such as those reported here) arises from an effect of F0 upon the auditory representation of vowel sounds. However, the data reported here suggest, that in addition to any effect of the auditory system, F0 normalization also includes a component which can only be described as an adjustment-to-talker.

References

Ainsworth, W. (1975). Intrinsic and extrinsic factors in vowel judgements. In G. Fant & M. Tatham (Eds.), *Auditory analysis and perception of speech.* London: Academic Press.

Bladon, R., Henton, C., & Pickering, J. (1984). Towards an auditory theory of speaker normalization. *Language and Communication*, **4**, 59–69.

Bladon, R., & Lindblom, B. (1981). Modeling the judgement of vowel quality differences. *Journal of the Acoustical Society of America*, **69**, 1414–1422.

Crowder, R. (1981). The role of auditory memory in speech perception and discrimination. In T. Myers, J. Laver, & J. Anderson (Eds.), *The cognitive representation of speech.* New York, NY: North-Holland.

Fox, R. (1985). Auditory contrast and speaker quality variation in vowel perception. *Journal of the Acoustical Society of America*, **77**, 1552–1559.

Fujisaki, H., & Kawashima, T. (1968). The roles of pitch and higher formants in the perception of vowels. *IEEE-AU*, **16**, 73–77.

Gerstman, L. (1968). Classification of self-normalized vowels. *IEEE-AU*, **16**, 78–80.

Horri, Y., House, A., & Hughes, G. (1971). A masking noise with speech envelope characteristics for studying intelligibility. *Journal of the Acoustical Society of America*, **49**, 1849–1856.

Johnson, K. (1988a). *Processes of speaker normalization in vowel perception.* Unpublished doctoral thesis, Ohio State University.

Johnson, K. (1988b). Intonational context and F0 normalization. *Research on speech perception progress report no. 14.* Bloomington, IN: Indiana University.

Klatt, D. (1980). Software for a cascade/parallel formant synthesizer. *Journal of the Acoustical Society of America*, **67**, 971–995.

Ladefoged, P., & Broadbent, D. (1957). Information conveyed by vowels. *Journal of the Acoustical Society of America*, **29**, 98–104.

Lehiste, I., & Peterson, G. (1959). Vowel amplitude and phonemic stress in American English. *Journal of the Acoustical Society of America*, **31**, 428–435.

Miller, R. (1953). Auditory tests with synthetic vowels. *Journal of the Acoustical Society of America*, **25**, 114–121.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (in press). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America.*

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review*, **72**, 407–418.

Parducci, A. (1975). Contextual effects: A range-frequency analysis. In E.C. Carterette & M.P. Friedman, (Eds.), *Handbook of perception, volume II.* New York, NY: Academic Press.

Slawson, A. (1968). Vowel quality and musical timbre as functions of spectrum envelope and fundamental frequency. *Journal of the Acoustical Society of America*, **43**, 87–101.

Sussman, H. (1986). A neuronal model of vowel normalization and representation. *Brain and Language*, **28**, 12–23.

Syrdal, A., & Gopal, H. (1986). A perceptual model of vowel recognition based on the auditory representation of American English vowels. *Journal of the Acoustical Society of America*, **79**, 1086–1100.

Traunmüller, H. (1981). Perceptual dimension of openness in vowels. *Journal of the Acoustical Society of America*, **69**, 1465–1475.

On the External Evidence for y-Insertion in American English[1]

Stuart Davis

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research for this paper has been supported by NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN. I would like to thank Dan Dinnsen and Keith Johnson for discussion and comments on this paper.

Abstract

Halle & Mohanan (1985) argue that English has a rule of y-insertion that inserts [y] before a vowel that surfaces as [u] in such words as *beauty*, *feud*, and *pure*. Anderson (1988) argues against a rule of y-insertion based on speech error evidence from Shattuck-Hufnagel (1986). In this paper, it is argued that the external evidence (ie, evidence based on such phenomena as speech errors and language games) supports the y-insertion analysis for American English. First, the speech error evidence is reconsidered and it is shown that they are better analyzed under a y-insertion analysis. Afterwards, it is argued that only the y-insertion analysis is compatible with the relevant data from the language game Pig Latin.

On the External Evidence for y-Insertion in American English

Introduction

Halle & Mohanan (1985) argue that English has a rule of y-insertion that inserts [y] before a vowel that surfaces as [u] in such words as *beauty*, *feud*, and *pure*. Anderson (1988) contends that English has no rule of y-insertion. Rather, words like *beauty*, *feud*, and *pure* have /y/ (or /I̯/) in underlying representation where it is in close relationship with the following vowel.[1] (Anderson suggests that derivatively /I̯/ is also in close relationship with the preceding onset consonant; nonetheless, he still represents it as being in closer relationship with the following vowel.)

Anderson (1988) argues that the speech error evidence from Shattuck-Hufnagel (1986) supports his analysis of an underlying /I̯u/ diphthong in words like *beauty*, *feud* and *pure*, and that the speech error evidence is incompatible with an analysis like that of Halle & Mohanan which posits opaque underlying vowels and a rule of y-insertion. In this paper, however, it is argued that the external evidence supports a rule of y-insertion. First it is shown that the speech error evidence which Anderson considered is in fact compatible with a rule of y-insertion for American English. Afterwards, it is argued that only the y-insertion analysis is compatible with the relevant data from the language game Pig Latin. First, though, we consider in more detail the type of data to which the rule of y-insertion is supposed to apply.

The Data

The American English data in (1) are all examples of words in which [y] appears before an [u]. In the account of Anderson (1986,1988) all these words would contain an /I̯/ on-glide before the /u/ in the underlying representation. In an account like that of Halle & Mohanan (1985), a [y] is inserted by rule before a vowel that surfaces as [u].

(1) puny
bugle
fume
view
reputation
contribute
perfume
Cuba
vaccuum
argue
huge
humanity
music
Communism

---

[1] I will use the symbols /y/ and /I̯/ interchangeably. Technically, /y/ represents a sound that is part of the syllable onset while /I̯/ represents a sound that is an on-glide to the vowel and so would be part of the syllable nucleus. Normally, when referring to an account like Anderson's I use /I̯/ and when referring to a rule-based account I use /y/.

The data in (1) are all examples of words where there is a [y] after a noncoronal consonant (ie, a labial or a velar) and before [u]. Note that the [y] occurs regardless of whether it is in a syllable with primary stress, a syllable with secondary stress, or a syllable with no stress. Interestingly, in many varieties of American English, [y] can be found after coronal consonants (and before [u]) only in syllables that do not carry the main stress. Such words are shown in (2a). In (2b), representative words with stressed syllables containing a coronal consonant followed by [u] are shown. These words do not surface with [y].

(2)

a.	b.
volume	voluminous
continue	continuity
erudite	ruby
perpetual	perpetuity
residual	residue
Mathew	Methuselah

In an analysis of the American English data in (1) and (2) that incorporates a rule of y-insertion, there would seemingly be a rule like that in (3). (cor=coronal)

```
(3)  ∅ ---> y / <[+cor]> ___     u
                       <-stress>  (Condition: if [+cor]
                                   then [-stress])
```

The major point of criticism that can be leveled against the [y]-insertion rule analysis of the above data (and, indeed, Anderson (1988) does briefly mention this at the end of his squib) is that it forces a distinction between two different types of American English [u]: those that can have [y] inserted before it and those that cannot. While the data in (1) and (2) are all examples of forms that can undergo the [y]-insertion rule in (3), words like those in (4) never undergo [y]-insertion even though the environment for the rule is met in many of these words.

(4)
booty
poor
voodoo
roof
tooth
goober
moon
food
noon
soot
cool
hoof

Analyses that incorporate a rule of [y]-insertion normally posit that the vowel [u] in (1) and (2) is of a different character than the vowel [u] in (4). In Halle & Mohanan (1985), for example, the [u] in (1) and (2) are viewed as being derived from underlying back unrounded vowels, while the [u] in (4) is viewed as being derived from an underlying back rounded vowel. A more concrete approach like that of Anderson (1986,1988) avoids trying to make an abstract distinction between two types of [u]. Rather, the data in (1) and (2) would simply contain an underlying diphthong /I̲u/ and the words in (4) would have underlying /u/; there would be no [y]-insertion rule. So, for example, on Anderson's account the difference in a pair like *feud-food* would be that *feud* has the diphthong /I̲u/ underlyingly while *food* has /u/. In Halle & Mohanan's account neither word contains an underlying diphthong. Rather, *feud* has an underlying back unrounded vowel and *food* has an underlying back rounded vowel.

Although Anderson's (1986,1988) approach avoids having to distinguish two different types of [u] in American English, it does have at least two disadvantages that are not found with the y-insertion analysis. First, it fails to account for the limited and predictable distribution of the on-glide /I̲/ (ie, it occurs in the environment reflected by the rule in (3)–where [y] represents the same phone as [I̲]). And second, in order to account for alternations like *continue-continuity* and *volume-voluminous* shown in (2) a rule deleting /I̲/ before a stressed /u/ and after a coronal consonant would be required. No such rule is mentioned by Anderson (1986,1988). (Although one suspects that in an approach like that of Anderson's a word like *ruby* in (2b) as well as other words like it such as *Tuesday* and *suit* would have underlying representations without the on-glide /I̲/ since it fails to occur in any of the surface allomorphs. Nonetheless a rule would still be needed to account for alternations like *continue-continuity*.[2]

In summary, the language-internal evidence for a rule of y-insertion in American English is inconclusive. A y-insertion analysis seems to require a more abstract vowel system (with underlying back unrounded vowels) while a more concrete analysis is unable to account for the limited distribution of [I̲].[3]

We now turn to the external evidence that bears on y-insertion.

---

[2] Borowsky 1986 offers a very interesting analysis of y-insertion that is similar to that of Anderson (1986,1988) in that she also views data like that in (1) and (2) as having an underlying /I̲u/ sequence. However, in Borowsky's analysis, which is framed in a theory of CV-Phonology, the /I̲/ is underlyingly not associated to any X-slot. Only later (in level 2) does it acquire an X-slot which is then syllabified as part of the onset. One shortcoming of Borowsky's analysis is that it violates the principle of Prosodic Licensing (Ito 1986). This principle ensures that the output of each phonological cycle is exhaustively syllabified. Unlicensed segmental material is subject to Stray Erasure at the end of each cycle. Thus the unlicensed /I̲/ at the end of the first cycle would be expected to delete.

[3] Additional motivation for an abstract vowel analysis comes from vowel alternations reflected by such pairs as *assume-assumption* and *Lilliput-Lilliputian* as discussed by Halle & Mohanan (1985). Wang & Derwing (1986) suggest that this particular Vowel Shift alternation (ie, the alternation between [u] and [ʌ]) does not reflect a phonological rule based on spoken language alone but a spelling-sound correspondence rule. However, alternatively, one may consider this alternation a reflection of a (lexical) phonological rule that is reinforced by the orthography, at least among educated speakers.

Speech Errors

In this section we (re)consider the speech error evidence that bears on y-insertion. It is shown that contrary to Anderson (1988) the speech error evidence from Shattuck-Hufnagel (1986) is indeed compatible with the y-insertion analysis and is arguably superior to an analysis that posits /I̱u/.

Anderson (1988) contends that the speech error data from Shattuck-Hufnagel (1986) support the position that words like that in (1) have an underlying on-glide /I̱/ and that there is no rule of y-insertion. This is because some speech errors seemingly move /I̱u/ together as a unit while no speech errors seem to involve moving /I̱/ (or /y/) with the preceding consonant as a unit. Examples of errors that seem to move /I̱u/ as a unit which Anderson cites from Shattuck-Hufnagel (1986:131) are given in (5a) and (5b). Other similar type errors from Shattuck-Hufnagel are given in (5c)-(5e):

(5) a. piece (of music) ---> /pyu:s/
b. (new) mown (hay) ---> /myu:n/
c. occupied ---> occup/yu/
d. features (use) ---> futures
e immediate future ---> imm/yu/diate

While Anderson essentially interprets these errors as involving an anticipation (in 5a, 5d, and 5e) or a perseveration (in 5b and 5c) of an /I̱u/ sequence they are also quite compatible with an analysis where [y] is inserted by rule. Under a y-insertion analysis there is an anticipation or perseveration of the vowel which ends up in an environment where the y-insertion rule is met. For example, in *piece of music* an anticipation error occurs: the [u] of the word *music* is anticipated and so an [u] surfaces in the nucleus of the intended word *piece*. As a result of the anticipation the environment for the y-insertion rule is met, and it applies yielding [pyus]. Consequently, a y-insertion analysis is not incompatible with the speech error data in (5).

Although these two possible accounts for the speech errors in (5) are basically equivalent, the y-insertion analysis of errors like those shown in (6) is arguably superior. (The errors in (6a)-(6c) are errors that Anderson cites from Shattuck Hufnagel (1986:132). The errors in (6d)-(6f) are additional errors of the same type that are recorded in Shattuck-Hufnagel)

(6) a. feud ---> /flu:d/
b. (helicopter) crew ---> /kyu/
c. (marital) skew ---> /skru:/
d. cucumbers ---> cruk
e. fuse blown ---> fluz
f. future lay ---> flu

Anderson contends that in errors like those in (6), /I̲/ interacts with an onset consonant. Given Anderson's contention that /I̲/ is more strongly connected to the vowel than to the preceding consonant such an interaction between an element that is part of the nucleus and one that is part of the onset might be considered unusual. However, in all liklihood, these errors do not involve interactions between /I̲/ and an onset consonant, per se. Rather these errors are very similar to the ones in (5) in that they involve anticipations and perseverations, as well as a deletion in (6b). (Since Shattuck-Hufnagel did not provide the context for the error in (6a) it is unclear whether it should be classified as an anticipation error, a perseveration error, or an insertion error.) Consider the speech error in (6c) in which there is a perseveration error involving the phoneme /r/ from the word *marital*. When /r/ surfaces before the vowel in the word *skew*, the environment for y-insertion is no longer met (since there is no coronal before the vowel), so no [y] surfaces. Similarly, the errors in (6d)-(6f) all involve anticipations of a liquid. When /r/ or /l/ appear before the vowel that surfaces as [u], [y]-insertion is blocked from applying. Anticipation and perseveration errors of this sort are a very common type of speech error in English. But under Anderson's analysis the errors in (6c)-(6f) are not viewed as errors of this type.

The speech error in (6b) is revealing. If under Anderson's account the errors in (6) involve an interaction between /I̲/ and an onset consonant, then it is quite unclear where the /I̲/ comes from that interacts with the /r/ in (6b). In all liklihood, the error in (6b) is an example of a segment (ie, /r/) deletion possibly under the influence of the /r/ at the end of the preceding word *helicopter*. On an analysis that treats the [u] in data like in (1) and (2) as underlying /I̲u/, the vowel in a word like *crew* would have to be considered as underlying /u/ and not /I̲u/ since the /u/ of *crew* never alternates with [Iu].[4]

Nonetheless, when the speech error of deleting the [r] occurs a [y] appears. There is no apparent reason why on Anderson's account (or even Borowsky's account–see footnote 2) for [u] to be pronounced as [Iu] when the [r] of "crew" deletes in the error. On the account offered here, when the /r/ deletes in the speech error the environment for y-insertion is met (since the vowel no longer is preceded by a coronal), and the rule applies yielding [kyu:].

There are two other speech errors that Anderson (1988) mentions. These are:

(7) a. new (machines do) ---> /mju/
b. single unit ---> /sjunIt/

Anderson sees these as examples of errors leaving /I̲u/ intact. Under an analysis incorporating y-insertion, the first error could be considered an anticipation error with y-insertion applying between the /m/ and the vowel. The second error is classified as a blend by

---

[4] The same point is made by Borowsky (1986:294) with the example *rude*.

Shattuck-Hufnagel. Here, the entire second word remains intact.

What we have attempted to show in this section is that speech errors like those in (5), (6), and (7) are not incompatible with a y-insertion analysis. Thus, Anderson's observation that no speech errors seem to move [I̱] (or [y]) with the preceding consonant as a unit does not constitute evidence against a y-insertion analysis. His observation is only evidence against the presence of [y] as an underlying part of the onset in [Cyu]-sequences.[5]

Pig Latin

In this section it is argued that only an analysis of American English incorporating a rule of y-insertion can account for the relevant data from the language game Pig Latin.

In Pig Latin, words are formed by moving the initial consonant(s) of a word to the end and then adding the vowel [e]. In a transformational rule format the Pig Latin rule can be written as follows (where X represents a string of 0 or more phonemes after the first vowel):

```
(8)    # C  V  X ===> 1 3 4 2 e
       1 2  3  4
```

For example, the English words *lip* and *clip* would have the Pig Latin forms [Iple] and [Ipkle], respectively.

Words that phonetically begin with a [Cyu]-sequence are problematic. There are three common Pig Latin variants for these words. These are shown in (9) and are exemplified with the word *cute*.

```
(9)  a.  CyuX ===> uXCe     eg.  cute ---> [utke]
     b.  CyuX ===> uXCye    eg.  cute ---> [utkye]
     c.  CyuX ===> yuXCe    eg.  cute ---> [yutke]
```

The variants in (9a) and (9b) are incompatible with a view like that of Anderson in which /I̱u/ (or /yu/) is an underlying diphthong. If /I̱/ (or /y/) is in closer relation to the vowel /u/ than to the preceding consonant then it would be completely unexpected for /I̱/ to move with the consonant as in variant (9b) or for it to delete alltogether as in (9a). (9c) would be the only predicted variant.

---

[5] There is a single error that Shattuck-Hufnagel (1986:132) lists that seems to treat a Cy-sequence together. The error involves the intended utterance *redistribution* being mispronounced as *redistri/byei/*. This error would be very problematic under Anderson's account since it involves the splitting up of an underlying diphthong.

In an analysis that makes use of a y-insertion rule as in (3), the three variants are easily accounted for. In variants (9a) and (9b) the Pig Latin rule (8) applies after y-insertion. The only difference between (9a) and (9b) is whether or not speakers can violate the phonotactic restriction on Cye-sequences: those that can violate it have variant (9b), and those that cannot violate it have variant (9a).[6]

For variant (9c), the ordering of the rules are reversed: the Pig Latin rule occurs before the rule of y-insertion. In (10) the derivations for the Pig Latin variants in (9b) and (9c) are provided.

(10)		/kut/			/kut/
a.	y-insertion	kyut	b.	Pig Latin	utke
b.	Pig Latin	utkye	a.	y-insetion	yutke
	Surface Form	[utkye]		Surface Form	[yutke]

The Pig Latin variants thus can be taken as evidence that American English has a rule of y-insertion.[7]

In conclusion, the external evidence does not support an analysis of American English in which there is an underlying diphthong /I̯u/. The speech error data are compatible with the y-insertion analysis, and, moreover, Pig Latin variants like those in (9a) and (9b) are only compatible with an analysis incorporating a rule of y-insertion.

---

[6] Words in language games can sometimes violate the phonotactic restrictions of the corresponding language. See, for example, Kenstowicz & Kisserberth's (1979:167-168) discussion of the language game in the African language Sanga.

[7] Alternatively, for the variant [yutke], the rule ordering could still be y-insertion rule followed by the Pig Latin rule, if the Pig Latin rule for these speakers is modified so as to move only [+consonantal] sounds to the end of the word.

References

Anderson, J. (1986). Suprasegmental dependencies. In J. Durand (Ed.), *Dependency and non-linear phonology*, 55-133. London: Croom Helm.

Anderson, J. (1988). More on slips and syllable structure. *Phonology*, **5**, 157-159.

Borowsky, T. (1986). *Topics in the lexical phonology of English*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Halle, M. & Mohanan, K. (1985). Segmental phonology of Modern English. *Linguistic Inquiry*, **16**, 57-116.

Ito, J. (1986). *Syllable theory in prosodic phonology*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Kenstowicz, M. & Kisserberth, C. (1979). *Generative phonology: Description and theory*. New York: Academic Press.

Shattuck-Hufnagel, S. (1986). The representation of phonological information during speech production planning: evidence from vowel errors in spontaneous speech. *Phonology Yearbook*, **3**, 117-149.

Wang, H. & Derwing, B. (1986). More on English vowel shift: the back vowel question. *Phonology Yearbook*, **3**, 99-116.

RESEARCH ON SPEECH PERCEPTION
Progress Report No. 14 (1988)
Indiana University

Vowel Length and Closure Duration in Word-Medial VC Sequences[1]

Stuart Davis and W. Van Summers

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This research was supported, in part, by NIH Training Grant *NS-07134-10* and, in part, by NIH Research Grant *NS-12179-12* to Indiana University in Bloomington, IN.

Abstract

Previous research has shown that English vowel length varies depending on the voicing characteristic of the following consonant. For stop consonants, closure durations also vary as a function of consonantal voicing. Generally, vowel-stop sequences containing voiced consonants show longer vowel durations and shorter closure durations than similar sequences containing voiceless consonants. These previous studies have focussed on stressed vowels in monosyllabic or bisyllabic words. Very little research has examined the effects of postvocalic voicing on stressless vowels. In the present study, we examined the influence of postvocalic voicing on vowel and closure durations in VCV́ and V́CV sequences. Subjects produced sentence pairs containing target words contrasting in intervocalic consonantal voicing (e.g., adopt-atop, tabbing-tapping). Both stressed and stressless vowels tended to lengthen before voiced consonants. However, the vowel-lengthening effect was not as consistent for stressless vowels as for stressed vowels. Closure durations following stressed vowels were longer for voiceless stops than for voiced stops. However, voicing effects on closure duration were inconsistent after stressless vowels. The results have implications concerning perceptual cues for intervocalic voicing, the syllabification of intervocalic consonants, and the nature of syllable-internal structure.

Vowel Length and Closure Duration in Word-Medial VC Sequences

Introduction

Many previous studies have shown that the length of English vowels will vary depending on whether the following consonant is voiced or voiceless. One of the earliest instrumental studies demonstrating this effects was that of Lehmann & Heffner (1943) who showed for English, "... that before a voiced stop one may expect a somewhat longer vowel duration than before the homorganic voiceless stop. This is true for every vowel, and our evidence on this point is unequivocal" [p. 43]. More recent research on English vowel durations have confirmed Lehmann & Heffner's original findings (House & Fairbanks, 1953; Peterson & Lehiste, 1960; Sharf, 1962; Chen, 1970; Luce & Charles-Luce, 1985, Summers, 1987).

Studies examining English consonant closure duration have shown that the closure period of an intervocalic voiced consonant is shorter than that of an intervocalic voiceless consonant (Lisker, 1957,1972,1978; Sharf, 1962; Weismer & Stathopoulos, 1983). Based on these studies of vowel duration and stop closure duration in English, it seems apparent (as Chen, 1970 has concluded) that there is temporal compensation (a negative correlation) between a vowel and a following consonant: A vowel is longer before a following (shorter) voiced consonant and is shorter before a following (longer) voiceless consonant.

Most of the above-cited studies have focused on the relationship between a stressed vowel and the following consonant in English. Previous work has generally examined stressed vowels in monosyllabic words (Peterson & Lehiste, 1960; Raphael, 1972) or stressed vowels in bisyllabic words (House & Fairbanks, 1953; Lisker, 1957, 1972, 1978; Sharf, 1962). In the literature discussing the effect of voicing on the duration of a preceding vowel, very little attention has been paid to stressless vowels. Do stressless vowels lengthen before (word-internal) voiced consonants just as stressed vowels do? Are closure durations following stressless vowels longer for voiceless stops than for voiced stops just as they are following stressed vowels? As far as we are aware the first question has not been addressed in the published literature for English. As for the second question, Lisker (1957,1972) states that the closure duration difference between a voiced and voiceless stop only occurs after a stressed vowel. However, no actual data is presented showing this. Stathopoulos & Weismer (1983) state that intervocalic voiceless closures are always longer than intervocalic voiced closures regardless of whether the preceding vowel is stressed or stressless. However, judging from their data, the difference in duration after a stressless vowel is probably not statistically significant since the voiceless closure is only four to six milliseconds longer than the corresponding voiced closure.[1] They do not give any statistical tests.

The present research examines whether or not the negative correlation between the duration of a vowel and the duration of the following consonantal closure holds for unstressed vowels. We examine unstressed vowels by focusing on polysyllabic words since in monosyllabic (content) words, vowels are always considered stressed. Specifically, we consider whether the unstressed vowel in a word-internal VCV́ sequence lengthens before a voiced consonant and whether the closure duration of the intervocalic consonant in such sequences is longer for voiceless consonants than for voiced ones (as is the case after a stressed vowel). We also examine whether voicing effects are similar on stressless vowels that are reduced to schwa (ə) and on stressless vowels that are not reduced. The effects of postvocalic voicing on vocalic and consonantal durations for these two types of word-internal VCV́ sequences are compared to voicing effects within the more frequently examined word-internal V́CV sequences.

Method

Subjects. Three male native English speakers (KJ, CM, and MC) participated as subjects. Each speaker was a member of the laboratory staff who participated voluntarily. Speakers were naive to the purpose of the study and did not report any hearing or speech problem at the time of testing.

Stimulus Materials. Stimulus materials consisted of the 9 sentence pairs (18 sentences) listed in Appendix A. Each sentence was composed of a sentence frame and a word containing a target vowel-consonant-vowel (VCV) sequence. Sentence frames were phonetically similar for a given sentence pair. Target words in sentences from a given sentence pair differed only in terms of the voicing feature of an intervocalic consonant (e.g., degrees/decrees).[2] Three types of target VCV sequences were represented in the nine sentence pairs. Three sentence pairs contained V́CV target sequences with a stressed vowel preceding the intervocalic consonant, three sentence pairs contained VCV́ sequences with an unstressed unreduced vowel preceding the consonant, and three pairs contained əCV́ sequences with an unstressed ə vowel preceding the consonant. The three sentence pairs representing a given VCV structure differed in terms of place of articulation of the intervocalic stop. One sentence pair contained labial stops (e.g., tabbing/tapping), one pair contained alveolar stops (e.g., sided/sighted), and one pair contained velar stops (e.g., sagging/sacking).

Procedure. Subjects were run individually in a single-walled sound-attenuated booth (IAC Model 401A). Subjects were seated comfortably in front of a CRT display and fitted with a headset containing an Electrovoice condenser microphone (Model C090). The microphone was attached to the headset with an adjustable boom. Once adjusted, the microphone remained at a fixed distance of 4 inches from the speaker's lips throughout the experiment.

The experiment consisted of 10-12 blocks of trials with 12 trials per block.[3] Each trial of the experiment consisted of three parts. In the first part, both sentences of a given sentence pair were shown on the CRT display. When the subject was familiar with the sentences, he made a button response to proceed to the second part of the trial. Here, one of the two sentences just shown was redisplayed. The subject produced the sentence aloud and made another button response to proceed to the third portion of the trial. Here, the other member of the sentence pair was displayed. The subject produced this sentence and made a button response to complete the trial. In each block of trials, the 12 pairs of sentences were presented in a randomized order.

A modified version of the Speech Research Laboratory's Speech Acquisition Program (Dedina, 1987) was used to digitize subjects productions during the experimental sessions. The utterances were low-pass filtered at 4.8 kHz and then sampled by a VAX 11/750 computer at a rate of 10 kHz using a 16-bit A/D converter.

Vowel, closure, and aspiration duration for the initial vowel and following intervocalic consonant of each target utterance were determined by visual inspection of a computer-graphics display of the utterance waveform and by simultaneously listening to tentatively marked segments. Following identification of segment boundaries, the duration of each marked segment was stored for further analysis.

Results and Discussion

In general, we found that word-internal vowels lengthen before voiced stops regardless whether the vowel is stressed or stressless. However, the difference was not always statistically significant for stressless vowels. Furthermore, while closure duration after stressed vowels was always longer for voiceless stops than voiced ones, it varied considerably after stressless vowels. The results for the three VCV sequences (V́CV, əCV́, and VCV́) are presented in the following sections.

V́CV Sequences

Results concerning the duration of a stressed vowel and following closure in a V́CV sequence are presented in Table 1. Mean vowel durations were greater before /b/ and /g/ than before their voiceless counterparts. The differences in mean vowel duration ranged from 12 ms to 23 ms and were statistically significant at these two places of articulation for each speaker. Closure durations were shorter for /b/ and /g/ than for their voiceless counterparts /p/ and /k/, respectively. Differences in mean closure duration ranged from 10 ms to 23 ms. These differences were also significant at each of these two places of articulation for each speaker. Subjects varied on whether they lengthened the vowel before /d/. Two subjects had significantly longer vowels before /d/ than before /t/. The third subject did not demonstrate significant lengthening. This pattern may reflect a difference among

Table 1. Vowel, closure, and aspiration durations for V́C sequences.

		voiced consonant		voiceless consonant		voiced mean -
vowel duration (ms)	spkr	mean	S.D.	mean	S.D.	voiceless mean
labial	KJ	110	7.7	98	9.9	12**
	MC	124	12.5	109	5.0	15*
	CM	151	9.2	134	7.9	17**
alveolar	KJ	139	25.0	130	21.6	9
	MC	169	15.8	126	11.1	43**
	CM	174	12.0	144	8.7	30**
velar	KJ	132	11.8	113	9.5	19**
	MC	167	20.1	144	11.8	23**
	CM	164	7.8	141	5.4	23**
closure duration (ms)						
labial	KJ	50	14.7	60	5.9	-10*
	MC	50	11.2	73	5.9	-23**
	CM	53	6.3	74	4.4	-21**
alveolar	KJ	31	13.3	27	19.0	4
	MC	34	4.4	36	4.7	-2
	CM	40	7.1	37	6.5	3
velar	KJ	31	16.9	44	12.2	-13**
	MC	40	10.8	60	18.9	-20**
	CM	44	5.7	66	8.4	-22**
aspiration duration (ms)						
labial	KJ	1	1.6	34	6.8	-33**
	MC	0	0.0	15	1.8	-15^{4}
	CM	5	3.7	20	4.7	-15**
alveolar	KJ	0	0.6	2	5.7	-2
	MC	8	6.6	18	7.3	-10**
	CM	8	5.4	14	9.4	-6
velar	KJ	0	0.0	48	19.4	-48**
	MC	1	3.6	23	24.3	-22*
	CM	21	6.4	41	8.4	-20**

* $p < .05$
** $p < .01$

American English speakers as to whether they implement vowel lengthening based on the underlying phoneme (/t/ or /d/) or based on the voiced flap allophone of the alveolar stop consonants. The former strategy would be expected to produced vowel duration differences, the latter would not.[5] Closure duration for intervocalic post-stress /t/ and /d/ varied little, most likely because these phonemes are neutralized in this position.

These results are in agreement with previous studies. Sharf (1962) and Lisker (1978) have found a negative correlation between the length of a stressed vowel and the closure duration of a following stop consonant in a word-internal V́CV sequence. Stressed vowels were longer before a (shorter) voiced closure than before a (longer) voiceless closure.[6] Lisker (1978) examined intervocalic labials in V́CV sequences. He found that stressed vowels were longer before /b/ than /p/ while closure durations were longer for /p/ than /b/. Sharf (1962) examined intervocalic labials, alveolars, and velars after stressed vowels. He found that stressed vowels were longer before /b/ and /g/ than before their voiceless counterparts by about 33 ms. He also found that the closure durations for /b/ and /g/ were on the average about 25 ms shorter than for their voiceless counterparts. The alveolar consonants, on the other hand, did not show this relationship. While, on the average, the vowel before /d/ was slightly longer than before /t/ (about 9 ms), the closure duration for /d/ was slightly longer than for /t/ (about 4 ms). The similar closure durations reported for intervocalic alveolar stops in V́CV sequences is actually not surprising since /t/ and /d/ are both pronounced as a voiced flap when in intervocalic position after a stressed vowel.

We also examined the duration of aspiration after the release of the stop closure. For all our subjects, aspiration durations for bilabial and velar stops following voiceless closure were significantly longer than aspiration durations following voiced closure. For alveolar stops, aspiration durations also tended to be longer for voiceless consonants than voiced consonants although these differences in aspiration duration were not always statistically significant.

In summary, our findings on vowel length and closure duration in word-internal V́CV sequences essentially replicate those of Sharf (1962) and Lisker (1978). There is a negative correlation that holds between the duration of a stressed vowel and the length of the closure of a following labial or velar consonant in a V́CV sequence. The present data and results reported by Sharf (1962) suggest that this pattern does not hold for intervocalic alveolar stops. These are generally produced as voiced flaps regardless of their underlying voicing feature so that differences in closure and aspiration duration are reduced or eliminated. Voicing effects on closure and aspiration duration for alveolar stops were generally small and nonsignificant in the present data.

əCV́ Sequences

In most instances, our three subjects produced the reduced vowel schwa (ə) with a longer duration before a voiced consonant than before its voiceless counterpart (see Table 2). These

Table 2. Vowel, closure, and aspiration durations for əC sequences.

		voiced consonant		voiceless consonant		
vowel duration (ms)	spkr	mean	S.D.	mean	S.D.	voiced mean - voiceless mean
labial	KJ	36	5.8	37	6.3	-1
	MC	77	8.4	72	5.0	5
	CM	45	5.6	38	3.3	7**
alveolar	KJ	32	8.9	26	6.1	6
	MC	74	14.1	73	14.8	1
	CM	29	12.9	33	12.9	-4
velar	KJ	29	5.3	21	5.5	8**
	MC	55	13.9	49	5.1	6
	CM	42	5.3	34	4.4	8**
closure duration (ms)						
labial	KJ	67	6.6	67	5.4	0
	MC	89	5.6	76	3.5	13**
	CM	101	9.8	101	7.6	0
alveolar	KJ	63	6.0	67	4.3	-4**
	MC	83	14.8	54	8.1	29**
	CM	95	5.0	82	7.4	13**
velar	KJ	46	6.0	55	5.2	-9**
	MC	68	14.5	59	8.2	9
	CM	56	8.6	63	7.5	-7**
aspiration duration (ms)						
labial	KJ	18	6.5	49	6.2	-31*
	MC	0	0.0	70	7.4	-70^4
	CM	16	7.0	67	21.7	-51**
alveolar	KJ	25	4.6	57	6.0	-32
	MC	17	4.1	89	6.5	-72
	CM	16	1.9	71	12.2	-55
velar	KJ	40	12.0	78	6.7	-38**
	MC	32	5.3	105	10.4	-73**
	CM	47	13.9	95	6.7	-48**

* $p < .05$
** $p < .01$

durational differences were generally quite small and there were exceptions to this general pattern. KJ's productions of *abase* and *apace* and CM's productions of *adopt* and *atop* showed slightly longer durations for vowels preceding voiceless consonants than voiced consonants. In all other cases, schwa vowels preceding voiced consonants had longer durations than those preceding voiceless consonants.

All cases in which vowel duration differences were statistically significant were in the direction of longer vowel durations for vowels preceding voiced consonants. Specifically, for two of the three subjects, the difference in vowel duration was statistically significant for the pair decree/degree with the vowel being longer in *degree* than *decree*. For the labial pair apace/abase, for one subject, the initial vowel was significantly longer in *abase* than *apace*. In general, schwa vowels tended to demonstrate longer durations before voiced consonants than before voiceless ones.

Consonantal voicing did not have a consistent influence on closure durations following schwa vowels. One of our subjects (MC) consistently produced longer closure durations for voiced stops after schwa. Another subject (KJ) tended to produce longer voiceless closures after schwa. The third subject (CM) did not show a consistent pattern. His closure durations for the /b/ and /p/ in the labial pair abase/apace were about equal. But he displayed longer voiced closures for the alveolar pair, adopt/atop, and longer voiceless closures for the velar pair, degree/decree. Thus, it appears that the length of the voiced or voiceless closure duration after a schwa is quite variable across speakers, unlike voiced/voiceless closure durations after a stressed vowel. Consequently, our findings do not support Stathopoulos & Weismer view that closure durations after stressless vowels are consistently longer for voiceless stops than for voiced stops.

While we found no consistent pattern across subjects in consonant closure durations after schwa, aspiration durations for stop consonants after schwa were quite consistent. All subjects produced voiceless stops with longer aspirations than voiced stops. Moreover, aspiration durations after schwa were almost always significantly longer than aspiration durations after stressed vowels. If aspiration duration is computed as part of the intervocalic consonant, durations of voiceless consonants are always significantly longer than durations of voiced consonants following schwa. This finding suggests that the duration of voiceless consonants in English are always longer than their voiced counterparts regardless whether the preceding vowel is stressed or reduced as long as aspiration is considered as part of the consonant.

VCV́ Sequences

In this section we report our findings concerning the relationship between an unstressed, unreduced vowel and an immediately following intervocalic stop. The relationship between an unstressed, unreduced vowel and a following stop was quite similar to what we reported in the previous section concerning the relationship between schwa and a following stop con-

Table 3. Vowel, closure, and aspiration durations for unstressed, unreduced VC sequences.

		voiced consonant		voiceless consonant		
vowel duration (ms)	spkr	mean	S.D.	mean	S.D.	voiced mean - voiceless mean
labial	KJ	70	7.0	74	5.2	-4
	MC	86	14.1	67	8.2	19**
	CM	69	10.6	69	8.3	0
alveolar	KJ	68	10.8	65	10.2	3
	MC	75	10.0	68	7.7	7
	CM	88	6.4	80	8.3	8*
velar	KJ	58	10.2	55	9.3	3
	MC	76	15.6	66	10.2	10
	CM	77	6.7	70	5.8	7*
closure duration (ms)						
labial	KJ	55	6.0	62	4.3	-7**
	MC	109	11.2	83	6.3	26**
	CM	102	10.8	95	8.7	7
alveolar	KJ	62	6.0	62	4.3	-7**
	MC	99	11.8	80	10.5	19**
	CM	86	14.3	74	12.7	12**
velar	KJ	60	7.3	58	10.6	2
	MC	74	9.2	65	11.8	9
	CM	50	6.4	52	9.3	-2
aspiration duration (ms)						
labial	KJ	3	5.3	41	12.8	-38**
	MC	14	4.0	91	12.5	-77**
	CM	15	8.5	67	7.7	-52**
alveolar	KJ	21	4.9	54	8.5	-34**
	MC	20	3.2	86	7.6	-66**
	CM	20	6.6	83	8.5	-63**
velar	KJ	36	8.4	72	12.4	-36**
	MC	32	10.5	109	9.2	-77**
	CM	44	7.1	97	4.8	-53**

* $p < .05$
** $p < .01$

sonant. In general, unreduced unstressed vowels tended to be slightly longer before voiced stops than before voiceless stops (see Table 3). These differences were not always significant. There was only one case where the opposite tendency was apparent. For the pair antibetting/antipetting, one of the subjects produced a longer vowel before the voiceless consonant. The difference in duration was not significant in this case.

Closure durations for intervocalic stops after stressless unreduced vowels were quite similar to those after schwa. The subject (MC) who had longer voiced closures after schwa also had longer voiced closures after unstressed, unreduced vowels. The subject (KJ) who tended to show longer voiceless closures after schwa also demonstrated longer voiceless closures after unstressed, unreduced vowels. The third subject (CM) who had longer voiced closures for alveolars, longer voiceless closures for velars, and about equal closure durations for the bilabial pair, had nearly the exact same pattern for closures after unstressed, unreduced vowels. Moreover, the aspiration duration of the stop consonants following an unstressed, unreduced vowel was quite similar to the aspiration data following schwa.

And, if aspiration is computed as part of the intervocalic consonant, the voiceless stops were always pronounced with greater duration than their voiced counterparts. Consequently, in our findings, there are virtually no differences between əCV́ sequences and other VCV́ sequences in terms of vowel durations, stop closure durations, and aspiration durations.

General Discussion

In this section we discuss the implications of our findings for three issues: the perceptual cues for voicing of intervocalic stop consonants, the syllabification of intervocalic consonants, and the nature of syllable-internal structure.

Voicing Cues in Intervocalic Stop Consonants

Two of the most important cues indicating voicing of a stop consonant that have been discussed in the literature are duration of the vowel immediately preceding the stop and the duration of the stop closure itself. The general finding is that a longer vowel duration is a cue that the following consonant is voiced while a longer closure duration for an intervocalic stop is a cue for voicelessness.[7]

Most of the work on cues for postvocalic consonantal voicing has focused on consonants immediately following a stressed vowel. This is seen in the work of Raphael (1972) who examined monosyllables and Lisker (1957) who looked at intervocalic consonants which immediately followed stressed vowels. Raphael (1972) found that with synthesized monosyllables, subjects perceived word-final consonants as voiceless when preceded by vowels of short duration and as voiced when preceded by vowels of long duration.[8] In a tape-splicing experiment, Lisker (1957) found that subjects perceived the intervocalic /p/ of *rupee* as a

/b/ when closure duration was reduced. Similarly, in another tape-splicing experiment in which the intervocalic /b/ in the word *ruby* was replaced by silence, subjects perceived the silence as /p/ when it was of relatively long duration and as /p/ when it was of shorter duration. Our findings are compatible with Raphael (1972) and Lisker (1957): our subjects consistently pronounced stressed vowels with longer durations before voiced stops than before voiceless stops. In addition, following stressed vowels, closure durations for voiceless consonants were longer than closure durations for voiced consonants.

However, our findings suggest that neither vowel duration nor closure duration are reliable perceptual cues for consonantal voicing when the preceding vowel is stressless. While our subjects tended to make stressless vowels slightly longer before voiced consonants than before the homorganic voiceless ones, the difference was not always statistically significant and in a couple of instances the vowel was actually longer before the voiceless consonant. Length of a stressless vowel, then, does not appear to be a consistently reliable cue to the voicing feature of a following stop consonant.

As for closure durations after stressless vowels, there was no consistent pattern across subjects. One subject had longer closure durations for voiced consonants after stressless vowels while another subject had longer closure durations for voiceless consonants. Our third subject tended to have longer closure durations for voiced bilabial and voiced alveolar stops but shorter closure durations for the voiced velar stop. These differences strongly suggest that closure durations after a stressless vowel are not a reliable cue for consonantal voicing. This finding argues against the proposal of Kluender et.al. (1988) who suggest that the closure duration difference between a voiceless and voiced stop is the main perceptual cue for intervocalic voicing. According to Kluender et al., the longer vowel duration difference that often occurs before a voiced stop provides an auditory enhancement of the closure duration cue. In their words:

> A longer preceding vowel makes a short closure interval appear even shorter, whereas a shorter vowel makes a long closure interval seem even longer. In other words, the contrastive effect of vowel duration augments the distinctiveness of the closure-duration cue. (p. 157)

In the present study, since closure durations following stressless vowels were not consistently longer for voicelss stops than voiced stops, closure duration could not be the main perceptual cue for intervocalic voicing. Therefore, the tendency for even stressless vowels to lengthen before voiced consonants could not be explained as auditory enhancement. If anything, our findings suggest that vowel length is a more reliable cue for intervocalic voicing than closure duration.[9] However, vowel duration was not an entirely reliable cue either.

What does seem to be a consistently reliable cue for the voicing characteristic of an intervocalic consonant is duration of the entire consonant. In our data, when aspiration was calculated as part of the voiceless consonant, the voiceless consonants had longer total duration (closure plus aspiration) than their voiced counterparts. This was true both when the voiceless consonant followed a stressless vowel and when it followed a stressed vowel. Total consonant duration thus seems to be a reliable perceptual cue for the voicing characteristic of an intervocalic stop. Stop closure duration and preconsonantal vowel duration were only consistently reliable cues for stop voicing in word-medial vowel-consonant sequences if the vowel was stressed.

Intervocalic Syllabification

In English, it is sometimes hard to determine what syllable an intervocalic consonant belongs to. It can be either a syllable-final consonant tautosyllablic with the preceding vowel or it can be a syllable-initial consonant tautosyllbic with the following vowel. Sometimes this can be determined based on the phonetics of the consonant. For example, the phoneme /l/ is normally pronounced as velarized (or "dark") when in syllable-final position and as "light" when in syllable-initial position. Also, and more relevant to our discussion, voiceless stops tend to be pronounced with strong aspiration when in syllable-initial position but not so otherwise. Given that strong aspiration of a stop correlates with syllable-initial position our data manifest a clear pattern concerning when an intervocalic stop should be considered syllable-initial or syllable-final. Voiceless stops are much more aspirated following a stressless vowel than following a stressed vowel. This supports the view that the intervocalic stop in such words as apace/abase, atop/adopt, and decree/degree is tautosyllabic with the following vowel. We suggest that this applies to voiced stops as well, even though they are only lightly aspirated: In the present data, the duration of aspiration of a voiced stop was consistently longer following a stressless vowel than a stressed vowel.[10]

On the other hand, our findings suggest that the intervocalic stops after a stressed vowel in such pairs as tapping/tabbing, sighted/sided, and sacking/sagging is tautosyllabic with the preceding stressed vowel because of their shorter aspiration durations. In other words, intervocalic stops after stressless vowels have aspiration characteristics of word-initial stops while the intervocalic stops after stressed vowels have the aspiration characteristics of word-final stops (which generally have shorter aspiration durations than word-initial stops). Thus, intervocalic stops after stressless vowels should be considered syllable-initial while intervocalic stops after stressed vowels should be considered syllable-final.

Syllable-Internal Structure

In the present data, English stressed vowels are longer before (shorter) voiced stops than before their (longer) voiceless counterparts. Stressless vowels also tended to be longer before voiced stops than before voiceless ones, although these differences are not always statistically significant. Researchers such as Selkirk (1982) and Walsh & Parker (1982) have taken the existence of the temporal relationship between the vowel and following consonant as evidence that the vowel and following consonant form a single unit or constituent, namely the rhyme constituent of the syllable. Selkirk (1982, p.353) states:

> Evidence such as that provided by Chen (1970), who claims that there is a constancy (aproximate) in the length of vowel plus stop combinations, could be taken as supporting the existence of the rhyme. According to Chen, a lengthening of the vowel (as before voiced stops) coincides with the shortening of the consonant. That is, one could say that within a constituent like the rhyme the duration of one element is adjusted in function of another.

Maddieson (1985), though, suggests that the lengthening of a vowel before a voiced stop is more akin to coarticulation phenomena, like anticipatory rounding, which ignores syllable and word boundaries. By this view, vowel lengthening before a voiced stop would not constitute evidence for syllable-internal structure, contrary to what Selkirk states.

Does the longer vowel duration before voiced consonants provide evidence for syllable-internal structure? For the view of Selkirk (1982) to be correct, it must be the case that vowel duration is longer before a voiced consonant than a voiceless consonant when both are in the same syllable (in a VC$ sequence, where $ represents a syllable boundary) but not when the vowel and consonant are in separate syllables (in a V$C sequence). The reasoning is as follows. If the occurrence of vowel lengthening is evidence for the constituency of a syllable-internal VC sequence, then vowel lengthening would be predicted not to occur over a syllable boundary since such lengthening would require that a V$C sequence forms a syllable-internal unit. On the other hand, if Maddieson's view is correct, it would be expected that the presence of a syllable boundary in a V$C sequence would have no influence on the length adjustment of the vowel due to the voicing characteristics of the following consonant. Maddieson describes the timing relationship between a vowel and a following consonant as a coarticulatory phenomenon. As such, it should not be sensitive to the presence of a syllable boundary. A vowel before a voiced stop would be lengthened whether they are tautosyllabic or heterosyllabic.

Both Selkirk and Maddieson correctly predict that the vowel lengthens before a voiced stop when the vowel and stop constitute a word-final VC-sequence (as, for example, in monosyllabic words). As already mentioned, Sekirk and Maddieson make different predictions on what would happen in a polysyllabic word when the intervocalic stop is a member of one syllable and the immediately preceding vowel is a member of the prior syllable. Selkirk predicts that there should be no vowel lengthening before a voiced stop since in such a case the stop would not be in the same syllable as the preceding vowel. Maddieson predicts that lengthening of the vowel would occur. According to Maddieson, the presence of the intervening syllable boundary should have no influence on a coarticulatory phenomenon like vowel lengthening before a voiced consonant.

At first glance, it would appear that Selkirk's contention cannot be correct because in the V́CV sequences of such word pairs as tapping/tabbing and sacking/sagging the stressed vowel is significantly longer before the intervocalic voiced consonant than before the voiceless counterpart. However, this would only be evidence against Selkirk's position if the syllabification of such sequences is V́$CV. Selkirk (1982), though, posits a (re)syllabification rule for English that syllabifies an intervocalic consonant before a stressless vowel with the syllable of the preceding vowel. That is, according to Selkirk, the sequence V́CV would be syllabified as V́C$V and not as V́$CV. Thus, the intervocalic stop in the pairs tapping/tabbing and sacking/sagging would be part of the first syllable. Given her syllabification rule, she correctly predicts that the vowel is lengthened when immediately before an intervocalic voiced stop in a V́CV sequence. Maddieson contends that vowels in general lengthen before a voiced stop regardless of syllabification. Therefore, he would concur with Selkirk's prediction in this instance.

The situation where Selkirk (1982) and Maddieson (1985) clearly make different predictions is when the vowel immediately following the intervocalic stop is stressed (i.e., in a VCV́ sequence). In Selkirk's theory of (English) syllabification, an intervocalic consonant preceding a stressed vowel forms a syllable with that vowel. Accordingly, a VCV́ sequence would be syllabified as V$CV́. Since in a V$CV́ sequence the stressless vowel and following consonant are heterosyllabic, Selkirk predicts that the stressless vowel would not lengthen before a voiced stop. On the other hand, Maddieson predicts that the stressless vowel would lengthen before a voiced stop. According to Maddieson, the syllable boundary plays no role in this relationship.

The data from our experiment tend to support Maddieson's view, but not conclusively. As we noted earlier, stressless vowels before intervocalilc stops tended to be longer before voiced stops than before the corresponding voiceless ones. The differences were not always statistically significant but the direction of the effect was fairly consistent. Since Selkirk would predict that there should not be any vowel lengthening before a voiced stop in a V$CV́ sequence, our findings do not support her view that the relationship between a vowel and the immediately following stop consonant provides evidence for syllable-internal constituency. Vowels tended to lengthen before a voiced stop regardless of whether they were in the same or different syllables. Thus, it appears that the relationship between vowel duration and the voicing feature of a following stop consonant does not provide conclusive evidence on the issue of English syllable-internal structure.[11]

Summary

In this paper we have examined the durational relationship between a vowel and a following stop in English VCV sequences. We found that the vowel durations were longer before intervocalic voiced stops than before their voiceless counterparts. This difference in vowel length was significant for stressed vowels, but was only a general tendency for stressless vowels. We also examined the intervocalic stop closure durations and aspiration durations. Closures were longer for voiceless stops when the preceding vowel was stressed. This pattern was not found when the preceding vowel was stressless. There seems to be individual variation as to whether or not voiceless closures are longer after stressless vowels. Thus a negative correlation between vowel length and stop closure duration is only found when the vowel is stressed. Regardless of stress, overall stop duration (closure plus aspiration) was consistently longer for voiceless stops than for voiced ones. These findings have implications for the issues of perceptual cues for intervocalic voicing, syllabification of intervocalic consonants, and the nature of syllable-internal structure.

Endnotes

[1] Crystal & House (1988) report durational data for unstressed vowels preceding voiced and voiceless consonants and for closure duration of these consonants. They did not find a negative correlation between vowel length and consonant closure duration. However, their data on stressless vowels involve, for the most part, V#CV sequences or VC#V sequences (where # is a word boundary). Virtually none of their data reflect word-internal VCV sequences.

[2] One of the pairs used was atop/adopt. Although *adopt* has an extra word-final consonant, this consonant /t/ was deleted in the sentence frame since it preceded the word *the*. Word-final post-consonantal /t/ almost always deletes in American English when the following word begins with a consonant. See, for example, Guy (1980).

[3] Subject MC participated in 10 blocks of trials while subjects KJ and CM participated in 12 blocks.

[4] No statistical test was conducted due to zero variance in one cell.

[5] See Fox & Terbeek (1977) for a specific study on English vowel duration before voiced flaps.

[6] Crystal & House (1988) do not report that closure durations were shorter for voiced consonants than for voiceless ones in VCV sequences. However, they report very little data on word-internal VCV sequences. In addition, the words they do have with these sequences were not all located in similar positions in the text sentences. For these reasons, we believe that their findings on closure durations in word-internal VCV sequences are not entirely conclusive.

[7] Vowel length is normally cited as a voicing cue for consonants in general, not just stop-consonants.

[8] Raphael does not address the issue of whether the vowel length cue for voicing is a universal. However, Keating (1985) reports that the phenomenon of vowel lengthening before voiced consonants does not occur in some Slavic languages. Thus, vowel duration may not be a cue for voicing in these languages.

[9] Luce and Charles-Luce (1985) report vowel and closure durations for monosyllabic CVC's which also suggest that vowel duration may be more reliable than closure duration as a cue to postvocalic voicing.

[10] The only exception to this pattern involved MC's productions of *abase* and *tabbing* which were consistently unaspirated.

[11] This seems to be true for other languages as well. See the general discussion in Davis (1985) and in Maddieson (1985). Kim (1975) presents a detailed study of vowel lengthening before stop consonants in Korean and concludes that the syllable boundary plays no role in the relationship.

References

Chen, M. (1970). Vowel length variation as a function of the consonant environment. *Phonetica*, **22**, 129-150.

Crystal, T. & House, A. (1988). Segmental durations in connected-speech signals: syllabic stress. *Journal of the Acoustical Society of America*, **83**, 1574-1585.

Davis, S. (1985). *Topics in syllable geometry*. Doctoral dissertation, University of Arizona, Tucson.

Dedina, M. (1987). SAP: A speech acquisition program for the SRL-VAX. In *Research on speech perception progress report no. 13*, 331-337. Bloomington, IN: Indiana University.

Fox, R. & Terbeek, D. (1977). Dental flaps, vowel duration and rule ordering in American English. *Journal of Phonetics*, **5**, 27-34.

Guy, G. (1980). Variation in the group and the individual: The case of final stop deletion. In W. Labov (Ed.), *Locating language in time and space*, (pp. 1-36). New York: Academic Press.

House, A. & Fairbanks, G. (1953). The influence of consonantal environment upon the secondary acoustical characteristics of vowels. *Journal of the Acoustical Society of America*, **25**, 105-113.

Keating, P. (1985). Universal phonetics and the organization of grammars. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged*, (pp. 115-132). Orlando: Academic Press.

Kim, K.-O. (1975). The nature of temporal relationship between adjacent segments in spoken Korean. *Phonetica*, **31**, 259-273.

Kluender, K., Diehl, R., & Wright, B. (1988). Vowel-length differences before voiced and voiceless consonants: an auditory explanation. *Journal of Phonetics*, **16**, 153-169.

Lehmann, W. & Heffner, R.-M. (1943). Notes on the length of vowels. *American Speech*, **18**, 208-215.

Lisker, L. (1957). Closure duration and the intervocalic voiced-voiceless distinction in English. *Language*, **33**, 42-49.

Lisker, L. (1972). Stop duration and voicing in English. In A. Valdman (Ed.), *Papers in linguistics and phonetics to the memory of Pierre Delattre*. The Hague: Mouton, 339-343.

Lisker, L. (1978). Segment duration, voicing and the syllable. *Status report on speech research*, **SR-54**, 175-189.

Luce, P.A. & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *Journal of the Acoustical Society of America*, **78**, 1949-1957.

Maddieson, I. (1985). Phonetic cues to syllabification. In V. Fromkin (Ed.), *Phonetic linguistics: Essays in honor of Peter Ladefoged*. Orlando: Academic Press, 203-221.

Peterson, G. & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America*, **32**, 693-703.

Raphael, L. (1972). Preceding vowel duration as a cue to the perception of the voicing characteristics of word-final consonants in American English. *Journal of the Acoustical Society of America*, **51**, 1296-1303.

Selkirk, E. (1982). The syllable. In H. van der Hulst & N. Smith (Eds.), *The structure of phonological representations part II*, (pp. 337-383). Dordrecht: Foris Publications.

Sharf, D. (1962). Duration of post-stress intervocalic stops and preceding vowels. *Language and Speech*, **5**, 26-30.

Stathopoulos, E. & Weismer, G. (1983). Closure duration of stop consonants. *Journal of Phonetics*, **11**, 395-400.

Summers, W.V. (1987). Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. *Journal of the Acoustical Society of America*, **82**, 847-863.

Walsh, T. and Parker, F. (1982). Consonant cluster abbreviation: An abstract analysis. *Journal of Phonetics*, **10**, 423-438.

Appendix: Stimulus sentence pairs.

V́CV sequences

They're tabbing him for the interview.
They're tapping him for the interview.

The sagging of the economy ended their hopes.
The sacking of the quarterback ended their hopes.

He often sided with their enemy.
He often sighted with their telescope.

əCV́ sequences

He will not abase the leaders.
He's still not apace the leaders.

He saw the group adopt the platform.
He saw the group atop the platform.

Various decrees were given out by the king.
Various degrees were given out by the dean.

VCV́ sequences

The anti-petting laws are unpopular.
The anti-betting laws are unpopular.

The anti-teen faction ran the high school.
The anti-dean faction ran the college.

The anti-crane lobby wanted the birds to die.
The anti-grain lobby wanted the wheat to rot.

Detailing the Nature of Talker Normalization in Speech Perception[1]

John W. Mullennix and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] The research reported here was supported, in part, by NIH Research Grant *NS-12179-12* and, in part, by NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN. The authors would like to thank Amy Lawlor for assistance in running subjects and Luis Hernandez for programming assistance.

Abstract

Previous studies have shown that changes in a talker's voice from stimulus to stimulus affect the perception and recall of spoken words (see Creelman, 1957; Mullennix et al., in press; Martin et al., in press). In the present study, the effects of talker variability were investigated in further detail. Performance in an AX same-different task was compared across three talker conditions: single talker, multiple talker, and multiple mixed talker. In addition, the effects of talker variability over time were examined by observing performance at different ISI delays. The results showed that response latencies were slower as a function of trial-to-trial and word-to-word talker variability. However, the effects of talker variability did not change over time. The results suggest that information about a talker's voice is maintained in memory and appears to be involved in the processes associated with developing phonetic representations of the spoken input.

Detailing the Nature of Talker Normalization in Speech Perception

The results from a recent study suggest that the mechanism or set of mechanisms that adjust for differences in a talker's voice in speech perception have a processing "cost" associated with them (Mullennix, Pisoni, & Martin, in press). This processing "cost" has been documented in terms of effects on spoken word recognition (Mullennix et al., in press) and effects on memory for spoken words (Martin, Mullennix, Pisoni, & Summers, in press). In these studies, the perceptual consequences of talker normalization have been examined in various tasks by manipulating the amount of talker variability in the stimulus input. For example, Mullennix et al. (in press examined word recognition performance under conditions where words were produced by a single talker or by different talkers. Their results showed that word recognition performance was worse when the voice of the talker varied from trial to trial, indicating that the operation of talker normalization processes has detrimental effects on the perceptual processes involved in speech perception and spoken word recognition.

Talker variability also affects the processes involved in the encoding and/or rehearsal of spoken words and their transfer into long-term memory (Goldinger, Logan, & Pisoni, 1988; Martin et al., in press). Under certain conditions, recall for early-list items is worse when the words within the lists are produced by different talkers compared to words in lists produced by a single talker. Thus, there appears to be less time or capacity available for rehearsal processes when the words in the list are produced by a variety of talkers. Taken together, these studies suggest that talker normalization involves a resource-demanding mechanism that, due to its operation, directly or indirectly affects the processes involved in recognition and recall of spoken words.

These results constitute an important first step towards characterizing the nature of perceptual normalization processes in speech perception and their relationship to other cognitive processes. However, a number of important issues concerning the nature of normalization in speech perception remain to be investigated. One hypothesis that could be entertained concerning perceptual adjustments to the talker is that attributes of a talker's voice are maintained in memory and compared to the incoming perceptual correlates of new items. If the perceptual information matches the information already in memory, then a perceptual adjustment to the voice of the talker may not be necessary. However, if voice information doesn't match, then some perceptual adjustment may be mandatory. This description of the normalization process assumes that information about a talker's voice is held in memory for a period of time to permit comparison with incoming perceptual information.

There is some evidence from previous research that is consistent with the idea that information about a talker's voice is maintained in memory for a short period of time. Cole, Coltheart, and Allard (1974) and Allard and Henderson (1976) found that in a same-different matching task response latencies for "same" responses to consonant-vowel syllables and isolated vowels were slower when the stimuli in each trial differed in the voice of the talker.

In addition, these effects persisted up to periods of 8 seconds. Cole et al. interpreted their results in terms of the superiority of physical matches versus name matches. Thus, the mismatch between talker information in memory and incoming talker information in the next stimulus may have led to the engagement of talker normalization processes that incur a processing cost and affect with performance. This interpretation is bolstered by the fact that the effects lasted for 8 seconds, suggesting that information about a talker's voice was available in memory for a substantial amount of time.

The purpose of the present study is to provide some new empirical evidence bearing on the issues concerning the relationship of talker normalization to memory. In order to accomplish this goal, we adopted the methodology of Cole et al. (1974) in order to examine the time course of talker variability effects. In addition, we wished to extend the previous findings reported by Mullennix et al. (in press) by assessing the degree to which talker variability effects vary within and across trials in an experiment. By examining talker variability effects in more detail and by assessing talker variability effects as a function of time, we hoped that evidence could be obtained indicating whether information about a talker's voice is maintained in memory and, if so, whether this information is related to the processes involved in talker normalization.

Given these goals, an experimental situation was designed to assess talker variability effects under these conditions. The experimental procedure consisted of an AX same-different matching task. The inter-stimulus-interval (ISI) was varied from 100 ms to 2 seconds. The use of the AX task allows control over the time elapsing between stimuli so that the effects of talker variability over time can be examined. The ISIs we used differed from those of Cole et al. (1974) because we wanted to examine performance over a more fine-grained time frame than what they had used (Cole et al. used ISIs of 500 ms, 2 sec, and 8 sec; we used 100 ms, 500 ms, 1 sec, and 2 sec). In this study, we were specifically interested in the time course of talker variability effects as they related to perceptual processing and memory rather than the long-term representation of talker information in memory.

Talker variability was manipulated by creating three talker conditions. In the single talker condition, all the stimuli were spoken by the same talker. In the multiple talker condition, the stimuli within each trial were spoken by the same talker, but the voice of the talker varied from trial to trial. Finally, in the multiple mixed talker condition, the voice of the talker varied both within and across each trial. The arrangement of stimuli in this manner allows a detailed assessment of talker variability effects in order to provide further information about the underlying basis of our earlier findings.

In our previous study (Mullennix et al., in press), the voice of the talker changed from word to word (and thus, trial to trial) in the multiple talker conditions. In the present experiment, the voice of the talker changed from trial to trial in the multiple talker condition, but changed from word to word *and* trial to trial in the multiple mixed talker condition. The arrangement of the talker conditions in this manner permits a situation to be created where any putative facilitory effects of information about a talker's voice in memory can be examined. In both the single talker and multiple talker conditions, the voice of the talker is the same within the trial. If talker information resides in memory for a period of time, then one might expect that performance would be facilitated in these conditions compared to the multiple mixed talker condition. However, if talker information is not extracted and held in memory for a period of time and talker variability effects are simply related to a change in stimulus parameters from trial to trial, then one would expect no difference in performance between the two multiple talker conditions. Thus, evidence regarding the hypothesis that information about a talker's voice is maintained in memory and related to normalization can be obtained.

The predictions for the experiment are as follows. First, if talker variability has perceptual consequences, performance in the multiple mixed talker condition should be worse than in the single talker condition. This result would be consistent with previous work indicating that changes in a talker's voice have detrimental effects on perception. Second, if information about a talker's voice is maintained in memory and is used to facilitate perception, then performance in both the single talker and multiple talker conditions should be better than the multiple mixed talker condition. Finally, if information about a talker's voice is maintained in memory and does not decay over periods of time up to two seconds, then the effects of talker variability should be present at all ISI's and no interaction of talker variability with ISI should occur. That is, the effects of talker variability should not diminish as a greater amount of time elapses between stimuli within a test trial.

Method

Subjects. One hundred and twenty undergraduate students enrolled in introductory psychology courses at Indiana University served as subjects. Each subject took part in one 1-hour session and received partial course credit for participating in the experiment. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimulus Materials. The stimuli consisted of 15 naturally spoken English words obtained from seven male and eight female talkers all of whom spoke with a midwestern dialect. The stimuli were English monosyllablic words selected from the corpus of words used in the Modified Rhyme Test (House et al., 1965). Each talker's utterances were recorded on audiotape in a sound-attenuated booth (IAC Model 401A) using an Electro-Voice Model D054 microphone and a Crown 800 series tape recorder. Each stimulus item was pronounced in citation format in unique randomized lists for each talker. The words were subsequently converted to digital form via a 12-bit analog-to-digital converter at a 10 kHz sampling rate and stored as digital files. The target words were digitally edited to produce the final experimental materials used in the study. RMS amplitude levels among words were digitally equated using a software package designed to modify digital waveforms.

Procedure. Two experimental factors were manipulated: Talker variability and inter-stimulus-interval (ISI). Talker variability was manipulated within subjects by presenting the stimuli in three different stimulus sets: Single talker, multiple talker, and multiple mixed talker. In the single talker set, all AX stimulus pairs were spoken by the same talker. In the multiple talker set, the stimuli in each AX pair were spoken by the same talker but the talker varied from trial to trial. And, in the multiple mixed talker set, the stimuli in each AX pair were spoken by different talkers and the talkers varied from trial to trial. ISI was manipulated between subjects by setting the time interval between offset of the "A" stimulus and onset of the "X" stimulus at 100 ms, 500 ms, 1 sec, and 2 sec, respectively.

The subjects were divided equally into groups and randomly assigned to the four ISI conditions. The experimental procedure used a "same-different" AX matching task. Subjects were required to decide whether the two words presented on each AX trial were the "same" or "different". They were instructed to base their decisions on word identity only. Within each of the three talker stimulus sets, 120 test trials occurred. In the single talker condition, the AX pairs were formed by drawing upon the utterances produced by a single male talker. In the multiple talker and multiple mixed talker conditions, the AX pairs were formed by drawing upon the utterances produced by seven male talkers and eight female talkers. Within each of the three sets, the assignment of stimuli to AX "same" and "different" pairs was randomized. Half of the trials consisted of "same" AX pairs and half the trials consisted of "different" AX pairs.

The stimuli were presented binaurally over matched and calibrated TDH-39 headphones to the subject at a listening level of 80 dB. Subjects were run in small groups in sound-treated booths containing headphones and two-button response boxes. Subjects were instructed to respond as quickly and as accurately as possible by pushing one of two buttons. A warning light was illuminated before the presentation of each stimulus. Presentation of each stimulus occurred three seconds after all subjects had made a response or three seconds after a 2-second response interval had elapsed. Subjects received no feedback during the experiment. A one-minute rest period was inserted after each stimulus test set. Stimulus-to-response button assignment was also counterbalanced across subjects; the order of talker conditions was counterbalanced across subjects by means of a Latin square design. Identification accuracy and response latencies were recorded for all trials. Responses over 2 seconds were eliminated from subsequent analysis. Response latencies were measured from the onset of the "X" stimulus. Stimulus presentation and data collection were controlled on-line by a PDP-11/34A computer.

Results

The data were analyzed in terms of overall percent identification errors and response latencies. For each subject, mean percent error and mean response latencies were calculated over each of the talker conditions at each ISI for "same" and "different" trials. Response latencies were analyzed for correct responses only.

Insert Table 1 about here

Table 1 displays mean response latencies and mean percent identification error collapsed over subjects for talker condition, ISI, and trial (same or different). A significant difference in "X" stimulus duration was observed across talker conditions (605 ms and 551 ms, respectively, for the single talker condition and the two multiple talker conditions, $p < .05$) and a significant correlation of "X" duration with response time was observed ($r=.09$, $p < .001$). Since the raw response latencies may have reflected a bias towards shorter or longer response latencies across conditions, the latency values reported in Table 1 were obtained by subtracting the duration of the "X" stimulus on each trial from the response latency obtained on that trial (as timed from "X" stimulus onset). Thus, any biases across conditions due to this factor were corrected.

Insert Figure 1 about here

Figure 1 displays the response latency data collapsed over subjects and trial for the talker conditions as a function of ISI. A three-way ANOVA was conducted on the latency data for the factors of talker condition, ISI, and trial. A significant main effect of talker condition was obtained $F(2,232) = 89.9$, $p < .001$. Newman-Keuls posthoc tests revealed that response latencies were significantly different between all three talker conditions (0.2 ms for the single talker condition, 56.4 ms for the multiple talker condition, and 113.7 ms for the multiple mixed talker condition).

Insert Figure 2 about here

Figure 2 displays the response latency data collapsed over subjects and ISI for the talker conditions as a function of trial. A significant main effect of trial was obtained $F(1,116) = 164.4$, $p < .001$. Response latencies were faster on "same" trials compared to "different"

Table 1

Mean response latencies (in msec) and mean percent identification error (in parentheses) collapsed over subjects for talker and ISI conditions as a function of trial.

		Talker Condition		
ISI	Trial	Single	Multiple	Multiple Mixed
100	Same	-64.0 (2.2)	4.7 (1.4)	63.0 (7.9)
	Different	7.9 (2.2)	64.3 (2.3)	92.4 (2.1)
500	Same	-68.4 (1.9)	7.4 (1.7)	85.0 (7.9)
	Different	5.3 (2.1)	58.6 (1.9)	100.5 (2.0)
1000	Same	-3.4 (1.8)	66.3 (2.2)	146.0 (9.5)
	Different	37.8 (1.3)	115.2 (1.9)	162.9 (1.8)
2000	Same	9.2 (2.1)	31.1 (2.3)	107.0 (8.6)
	Different	77.5 (3.3)	104.1 (2.9)	153.2 (3.3)
	Mean	0.2 (2.1)	56.4 (2.1)	113.7 (5.4)

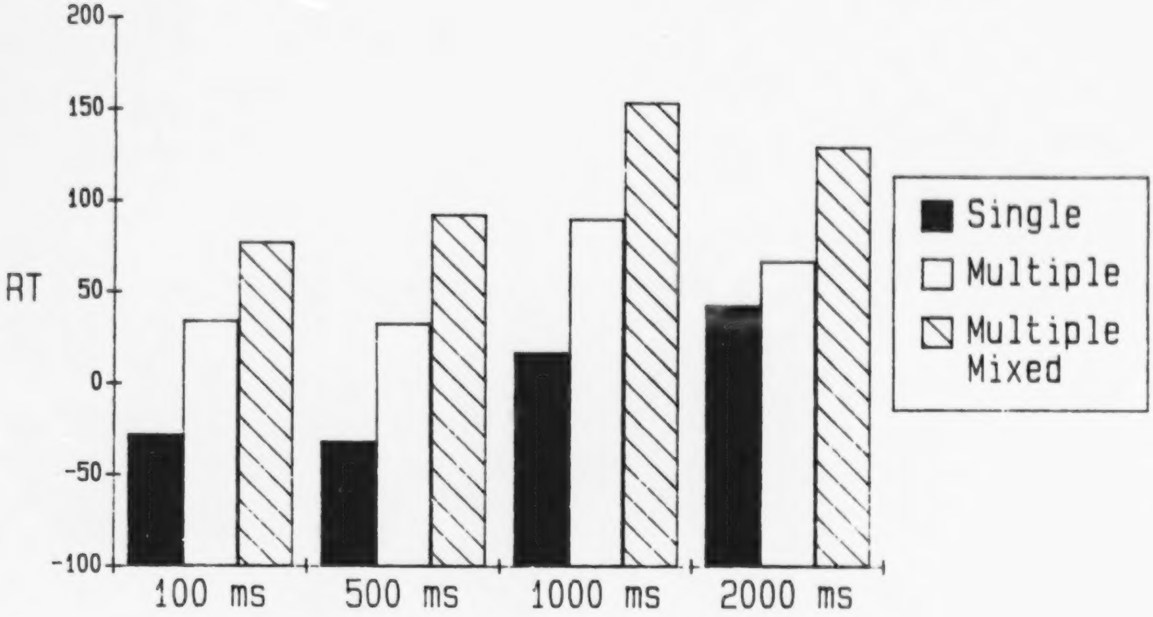


Figure 1. Mean response latencies collapsed over subjects and trial for talker condition and ISI.

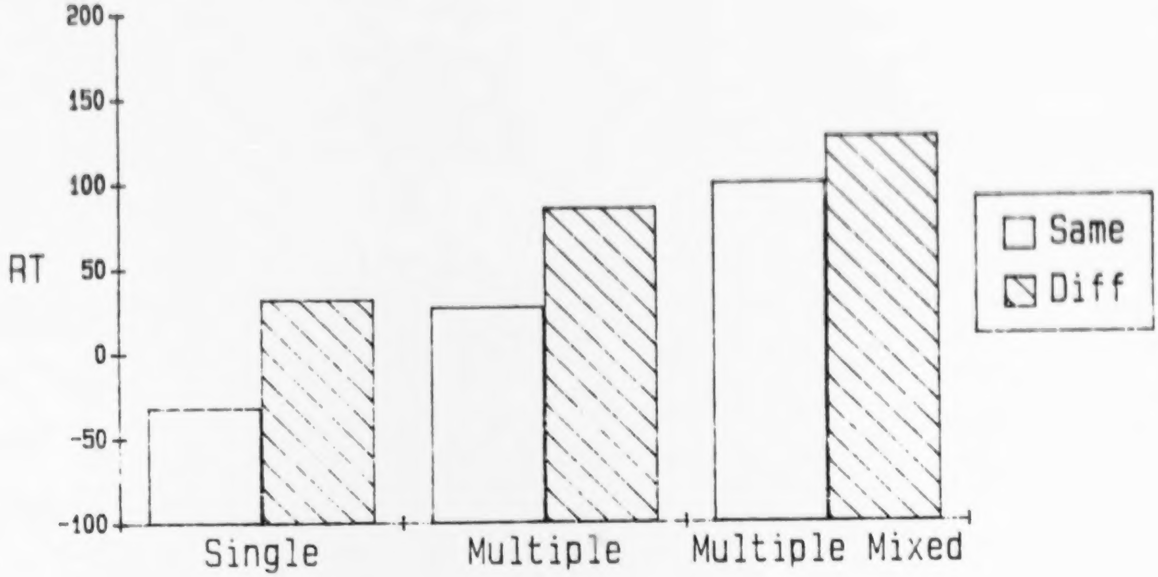


Figure 2. Mean response latencies collapsed over subjects and ISI for talker condition and trial.

trials (48.0 ms versus 81.6 ms). A significant interaction of talker condition with trial was also obtained $F(2,232) = 27.0$, $p < .001$). Post-hoc tests revealed that the difference between "same" and "different" trials was significant within each talker condition. No other significant differences between conditions were observed.

Insert Figure 3 about here

Figure 3 shows mean percent error collapsed over subjects and trials for the three talker conditions as a function of ISI. A three-way ANOVA was conducted on the arcsine transformed identification data. [1] A significant main effect of talker condition was obtained $F(2,232) = 136.8$, $p < .001$. Identification was more accurate in the single talker and multiple talker conditions and less accurate in the multiple mixed talker condition (1.9 %, 2.2 %, and 5.0 % errors, respectively). Post-hoc tests revealed that identification performance was significantly better in the single talker and multiple talker conditions compared to the multiple mixed talker condition. However, performance did not differ significantly between the single talker and multiple talker conditions.

Insert Figure 4 about here

Figure 4 shows the identification data collapsed over subjects and ISI for the talker conditions as a function of trial. A significant main effect of trial was obtained $F(1,116) = 86.4$, $p < .001$. Identification was less accurate for "same" trials compared to "different" trials. A significant interaction of talker condition with trial was also obtained $F(2,232) = 166.8$, $p < .001$). Post-hoc tests of the interaction revealed that the difference between "same" and "different" trials was significant only for the multiple mixed talker condition. A significant interaction of talker condition with ISI was also obtained $F(1,116) = 3.26$, $p < .03$). Post-hoc tests of this interaction revealed that the difference between "same" and "different" trials was significant for the 100 ms, 500 ms, and 1000 ms conditions, but not within the 2000 ms condition. No other significant differences between conditions were observed. Overall, in considering the error data and the latency data together, the pattern of results suggests that speed-accuracy tradeoffs do not account for the effects observed.

---

[1] All data analyses for the identification data were performed on nonlinear arcsine transformations of the raw percent correct identification data (see Cohen & Cohen, 1975).

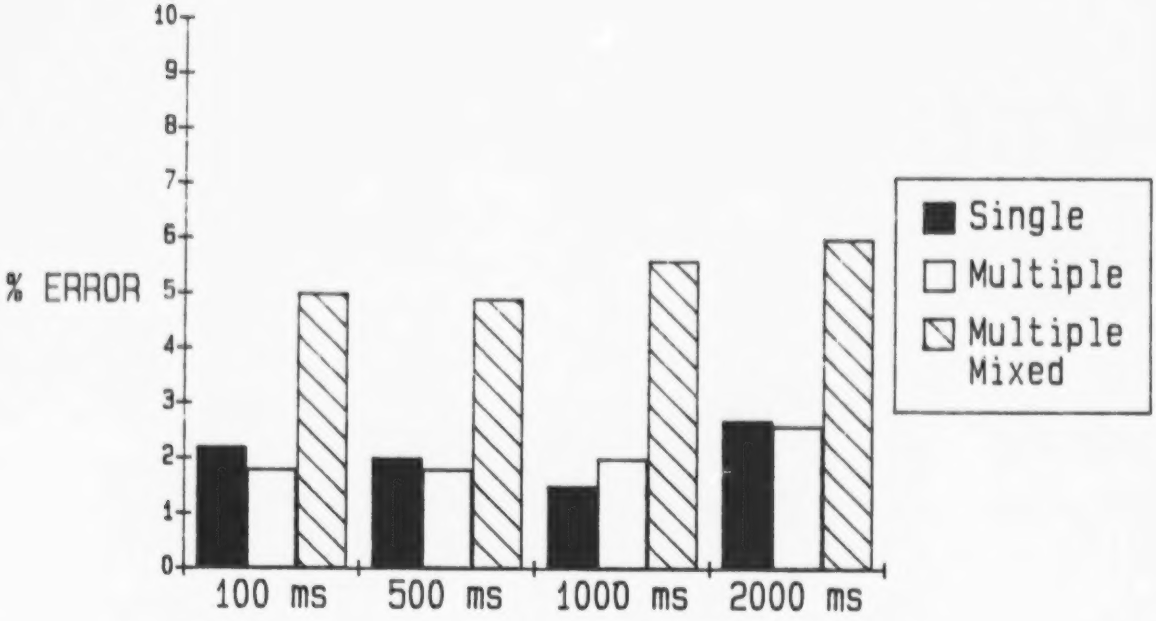


Figure 3. Mean percent error collapsed over subjects and trial for talker condition and ISI.

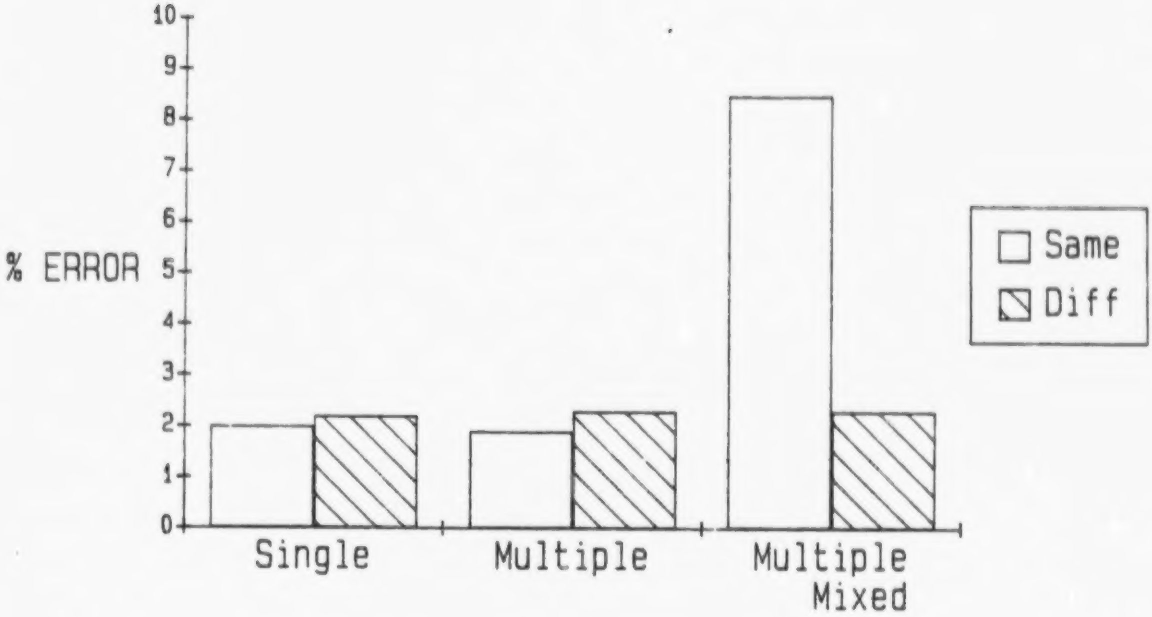


Figure 4. Mean percent error collapsed over subjects and ISI for talker condition and trial.

Discussion

The results of the present experiment provide further information about the effects of talker variability in speech perception. First, response latencies were faster for the single talker condition compared to the multiple mixed talker condition. This result indicates that perception of the stimulus items was affected when the items within each trial were spoken by different talkers. This result replicates our previous findings using perceptual identification and naming tasks (Mullennix et al. in press).

Second, response latencies were faster for the single talker condition compared to the multiple mixed talker condition at each ISI. In addition, we failed to find a significant interaction of talker condition with ISI. Taken together, these results indicate that the effects of talker variability did not significantly change over a time period of 2 sec. This pattern is consistent with the idea that talker information is extracted during perception, transferred to some representation in memory, and then maintained for a period of time. If talker variability effects had been found at short ISI's only, or if the talker variability effects had become smaller as ISI increased, the implication would have been that talker information is either lost during acoustic-phonetic processing or else is held in some fast-decaying sensory store. However, since these effects were still present 2 sec after offset of the stimulus item, the likely explanation is that talker information is, indeed, extracted from an utterance and then maintained in memory for periods of time sufficient for comparison to incoming perceptual information about the talker.

Third, response latencies were consistently faster for the multiple talker condition compared to the multiple mixed talker condition. Even though the voice of the talker changed from trial to trial in both conditions, when the information about a talker's voice within each trial was the same, perceptual performance was facilitated. This result indicates that it is not just the trial to trial changes in a talker's voice per se that leads to perceptual interference, rather, this result is consistent with the idea that information about a talker's voice is held in memory and is intimately related to talker normalization processes. Because performance was facilitated when the voice of the talker does not change, it is possible that talker normalization processes are not invoked or at least become more efficient when the information about a talker's voice in memory matches the perceptual talker information currently processed.

However, we also found that response latencies for the single talker condition were faster compared to the multiple talker condition. Thus, a degree of interference *was* caused by the trial to trial variation in a talker's voice when compared to the single talker condition. But, a possible explanation for this result is simply that, in the multiple talker condition, the information about a talker's voice extracted in a previous trial may have been present in memory at the time that the stimuli on the next trial were encountered. Since the talker voice always changed from trial to trial in this condition, once again a mismatch could have occurred between talker information in memory and perceptual talker information, leading to perceptual interference.

Finally, with regard to the pattern of identification errors, more errors were observed for "same" trials in the multiple mixed talker condition than for the other two conditions. Although we had no a priori reason to expect any differences across conditions in errors, it appears that the occurrence of more errors for "same" trials in the multiple mixed talker condition can be explained as a massive failure of selective attention to the acoustic-phonetic information specifying the word items. Since the voice of the talker varied between the two stimuli on each trial in the multiple mixed talker condition, subjects may have been distracted by the change in voice within the trial and responded "different" more often because the words were produced by different talkers. Thus, subjects' attention to word-related information was interfered with because the voice of the talker could not be selectively filtered out or ignored. In fact, the inability of subjects to selectively filter out information about a talker's voice was demonstrated by Mullennix and Pisoni (1987) using a selective attention procedure based on Garner (1974). Thus, the pattern of errors obtained in the present experiment appears to reflect an inability to selectively ignore information about a talker's voice. This result is consistent with the idea that selective attention must be allocated to information about a talker's voice in a mandatory fashion and cannot be selectively ignored in speech perception.

One final point concerning our present results is the question of the length of time talker information remains in memory and influences talker normalization. As mentioned earlier, Cole et al. (1974) found that at an 8 second delay "same" response latencies were still faster for same-talker versus different-talker trials. However, there is also evidence that at a 10 second delay the influence on performance of a previous talker is severely attenuated (Broadbent, Ladefoged, & Lawrence, 1956). On the basis of our results, we make no claims about how durable talker information in memory is. Our goal was only to assess performance at short ISI intervals in order to be able to closely track the time course of talker variability effects. The length of time that talker information remains in memory and the precise form of representation or representations that are involved have not been determined at this time.

It is also clear from the present results that talker normalization processes do not "normalize out" information about a talker's voice, leaving only a canonical acoustic-phonetic representation of the utterance that is passed on to higher-level processes (see also Goldinger, Logan, & Pisoni, 1988; Logan, Lively, & Pisoni, 1988; Martin et al., in press). Instead, information about a talker's voice remains in memory in some form and appears to be highly involved in phonetic categorization.

In summary, the results from the present study extend our earlier findings concerning talker variability effects in speech perception and provide further evidence concerning the relationship of talker normalization processes to memory. Information about a talker's voice appears to be extracted from spoken utterances and maintained in memory for at least a short period of time. One interpretation of this finding is that each time an utterance is encountered, talker information is processed and compared to talker information currently residing in memory. If the talker information matches, it is possible that the normalization mechanism is cued in some manner not carry out its cost-intensive operations. Alternatively, perhaps the perceptual adjustments to talker are always carried out, but by using the information about a talker's voice in memory in some manner the perceptual adjustments are carried out more efficiently. The present results cannot distinguish between these two alternatives. Overall, the results from this study represent an attempt towards extending our knowledge about the characteristics of perceptual normalization processes in speech perception and their relationship to the other processes in the cognitive system.

References

Allard, F., & Henderson (1976). Physical and name codes in auditory memory: The pursuit of an analogy. *Quarterly Journal of Experimental Psychology*, **28**, 475-482.

Broadbent, D.E., Ladefoged, P., & Lawrence, W. (1956). Vowel sounds and perceptual constancy. *Nature*, **178**, 815-816.

Cohen, J., & Cohen, P. (1975). *Applied multiple regression correlation analysis for the behavioral sciences.* Hillsdale, NJ: Erlbaum.

Cole, R.A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology*, **26**, 1-7.

Creelman, C.D. (1957). Case of the unknown talker. *Journal of the Acoustical Society of America*, **29**, 655.

Garner, W.R. (1974). *The processing of information and structure.* Hillsdale, NJ: Erlbaum.

Goldinger, S.D., Logan, J.S., & Pisoni, D.B. (1988). Determining the locus of talker variability effects on the recall of spoken word lists: Evidence from a presentation rate manipulation. *Research on speech perception progress report no. 14.* Bloomington, IN: Indiana University.

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

Logan, J.S., Lively, S.E., & Pisoni, D.B. (1988). Training Japanese listeners to identify /r/ and /l/: A first report. *Research on speech perception progress report no. 14.* Bloomington, IN: Indiana University.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (in press). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Mullennix, J.W., & Pisoni, D.B. (1987). Stimulus variability and processing dependencies in speech perception. *Research on speech perception progress report no. 13.* Bloomington, IN: Indiana University.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (in press). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America.*

Manner of Articulation and Feature Geometry: A Phonetic Perspective [1]

Keith Johnson

Speech Research Laboratory
Psychology Department
Indiana University
Bloomington, IN 47405

---

[1] Thanks to Stuart Davis and Ilse Lehiste for their comments and discussion on an earlier draft of this paper. Research supported by NIH Training Grant *NS-07134-10* to Indiana University in Bloomington, IN.

Abstract

Phonetic observations concerning the quantal nature of "degree of stricture" are used as the basis for a featural description of manner of articulation. The primary claims of the paper are: (1) that degree of stricture can be described in terms of a hierarchical organization of the three features [consonantal], [sonorant] and [continuant], (2) that this hierarchical structure constitutes the segmental "core" and (3) that other manner features such as [nasal] and [lateral] can be described as multiply articulated segments involving different degrees of strictures at different vocal tract locations.

Manner of Articulation and Feature Geometry: A Phonetic Perspective

Phonetic theory and phonological theory are converging upon a description of speech sounds. Approaches to feature geometry (Clements, 1985; Sagey, 1986b; McCarthy, 1988) are arriving at a description of "place of articulation" which is very much compatible with the traditional phonetic description (see Ladefoged and Halle, 1988). This convergence toward a unified description of speech sounds is both very exciting and incomplete. In particular, the description of "manner of articulation" still poses a problem. In this paper I offer some phonetic observations concerning manner of articulation, and speculate on the way these observations can be captured in a feature geometrical representation. The view of "manner of articulation" which I propose involves (1) primary reference to the notion "degree of stricture" and (2) combinations of different degrees of stricture and places of articulation for the description of the manner features [nasal] and [lateral].

Before addressing the problem of manner of articulation, however, it is instructive to consider the convergence of phonetic and phonological descriptions of place of articulation.

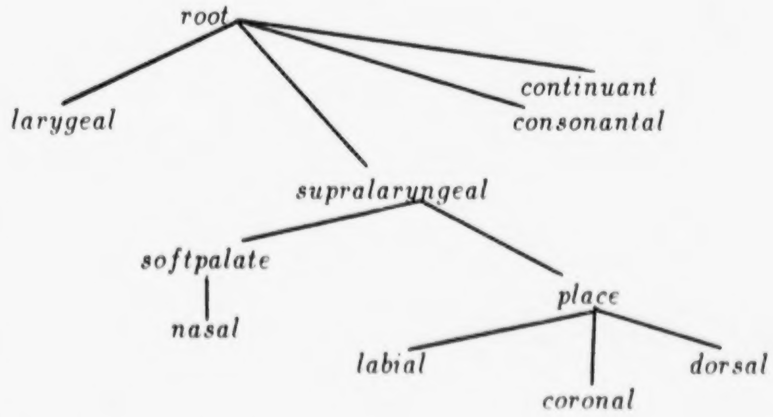
Convergence of Phonetic and Phonological Descriptions of "Place of Articulation"

Chomsky and Halle (1968) in the *Sound Pattern of English* (SPE) assumed that speech sounds are unstructured bundles of binary features. Place of articulation, in this view, is specified in terms of a number of logically unrelated features. So, for instance, the feature [anterior] divides the vocal tract (VT) at the alveolar ridge. All VT configurations which involve primary articulations at the alveolar ridge and forward are classified as [+anterior] and all VT configurations with primary articulation behind the alveolar ridge are classified as [-anterior]. The [+anterior] configurations are further subdivided by the feature [labial] and the [-anterior] configurations by the feature [coronal]. One difficulty of this scheme is that for [-anterior] configurations [labial] has no meaning.[1] If a segment is classified as [-anterior] there is no reason to specify any value for the feature [labial] (unless the segment is doubly articulated, e.g. [k͡p]). Likewise, the feature [coronal] is meaningless in the description of sounds made at the lips. So, there is an implicit dependency among the place features which is not captured in the SPE account of segments. Phonetic descriptions of place of articulation do not suffer from this overspecification because place is treated as a single, multivalent feature (Ladefoged, 1971, 1975). Features which relate to other places of articulation are not mentioned in the description of labials, rather, [labial] is simply one value which may be taken by the feature [place].

---

[1] Anderson (1976) found a clever use for [anterior] which hinged on its reference to the *primary* place of articulation. He proposed that doubly articulated stops such as [k͡p] be described as [+anterior] if the (phonologically determined) primary place of articulation is labial, and [-anterior] if the primary place of articulation is velar.

Recently, the conception of the segments as feature bundles has been revised in favor of a view in which dependency relations between features are captured in a hierarchical representation of features (Clements, 1985; Sagey, 1986b see (1)). In the feature geometrical representation, phonological theory is able to capture the insight that place of articulation is tied to the anatomy of the vocal tract, while at the same time avoiding the use of multivalued features. This is accomplished by positing the existence of privative features which correspond to the articulators. The presence of an articulator feature in the description of a sound indicates that the articulator is active during the production of that sound. The articulator features are not, however, defined for segments which do not involve the articulator to which the feature refers. Therefore, there is no set of speech sounds which can be characterized as [-labial], for example, as was possible with the SPE use of the feature. In addition, each of the articulator nodes dominates a set of features which are defined only for that articulator. Thus, [distributed] is only defined for sounds which involve the articulator [coronal] (the front part of the tongue). The structure of the geometry makes it impossible to specify any value for [distributed] unless the node [coronal] is also used in the description of the segment.



(1)

It is worth pointing out that I am arguing as a phonetician, rather than as a phonologist. I am focussing on the fact that the mention of the feature [labial] in the description of a speech sound, in the SPE approach, entails that the sound is [+anterior], and that, consequently, the specification of any value for [anterior] is redundant. This is a different conception of feature dependency than that expressed by Clements (1985:226), "If we find that certain sets of features consistently behave as a unit with respect to certain rules of assimilation or resequencing, we have good reason to suppose that they constitute a unit in phonological representation." In contrast to this criterion, I am focussing primarily on the description of speech sounds, and am arguing from what might be called "taxonomic dependencies". From the phonological perspective, on the other hand, the focus is upon the description of phonological processes and the arguments have to do with the economical statement of sound patterns. What is interesting is that the two viewpoints have led to as much convergence as they have. McCarthy (1988) compares the empirical consequences of the phonetic multivalent feature [place] and the feature geometry account of place of articulation and concludes that, "at the level of the gross architecture of featural structure, the two theories are not distinguishable from one another" (p. 14). The convergence of phonetic and phonological theories is indeed remarkable.

Manner of Articulation

The question of how to represent manner of articulation has not been answered with the unanimity accorded to the hierarchical approach to place of articulation. Clements (1985) proposed a manner node parallel to the place node in the geometry of features. However, phonological evidence for the existence of such a node was not forthcoming, so in a later paper (Clements, 1987) he dispenses with the "superfluous 'manner' node". Sagey (1986a) proposed that each articulator be populated by manner features, but later (1986b) retracted that view in favor of one in which the features [continuant] and [consonantal] attach directly to the root node of the segment (see (1) above). McCarthy (1988) proposes that the major class features [consonantal] and [sonorant] *are* the root node of the segment and that the manner features [continuant] and [nasal] are directly attached to the root. Ladefoged and Halle (1988) include a node similar to Clements' (1985) "manner" node which they call "stricture". The traditional phonetic approach to manner of articulation is illustrated by the presentation of the International Phonetic Alphabet (IPA). The symbols of the IPA for voiced and voiceless sounds are presented in a two dimensional matrix of which the dimensions are manner of articulation and place of articulation. The implicit analysis of manner features which is embodied in this table is one which is very similar to the feature geometries which place a manner node in the hierarchy parallel to the place node (Clements, 1985; Ladefoged and Halle, 1988).

The treatment of the manner features in a separate node, parallel to the place node (as in Ladefoged and Halle (1988) and in Clements (1985)) is initially attractive, but the fact that Clements' (1985) original proposal for manner features has not been widely accepted (compared with the general acceptance of his suggestion of a place node) indicates that parallel treatment for manner may be misguided. One difference between the manners of articulation and the places of articulation (both in the IPA presentation and in Clements' feature geometry) is that the manners of articulation are not a continuum along a single dimension as are the different places of articulation. However, among the IPA manners of articulation, there is a subset which can be viewed as steps along a continuum. The continuum is the traditional phonetic dimension "degrees of stricure" and includes plosive (stop), fricative, frictionless continuant (approximant), and vowel (resonant). If we limit our consideration of manner of articulation (at first) to this continuum, it may be possible to develop a coherent geometry of degree of stricture and, from this, a better understanding of the other manners of articulation. This, then, will be the outline of the discussion to follow. I first discuss the phonetic nature of the degree of stricture continuum. Then, I will propose a featural representation of the dimension and the location of the degree of stricture features within the segmental hierarchy. Following this, I will consider the manner features [nasal] and [lateral]. I will conclude the paper with a discussion of some remaining problems of a phonetic and phonological nature.

The Phonetic Description of Degree of Stricture. Stevens' (1972) observation concerning the degrees of stricture is that there are several "quantal regions" along the dimension. If we vary the degree of stricture at one particular place of articulation, for example at the lips, there are a limited number of possible acoustic effects for this essentially continuous articulatory dimension. The acoustic effect of complete closure is silence.[2] This is the first quantal region along the degrees of stricture dimension. When the closure is released slightly, and air is forced through the opening between the articulators, turbulence is created. In the case of constriction at the lips, the fricatives [ɸ] and [β] are produced, depending upon the state of the glottis. Further (slight) increase of the amount of opening will not change the fact that the airflow through the opening creates turbulence both in voiced and voiceless sounds. This region along the continuous articulatory dimension, degree of stricture, corresponds to a relatively stable acoustic output. Small changes in degree of stricture result in no appreciable change in acoustic output. However, there is a point at which further increase in the amount of opening will result in a configuration which does not create turbulence for voiced sounds, but does create turbulence in voiceless sounds. This definition of approximant was suggested by Catford (1977) and serves to point out the aerodynamic nature of the quantal regions of degree of stricture.[3] The reason for the difference between voiced and voiceless sounds is that the rate of airflow for voiceless sounds is greater than for voiced sounds because the glottal constriction involved in producing voicing serves as an obstruction in the airstream. Again, there is a range of degrees of stricture for which these conditions hold and the acoustic output is relatively stable. Then, at some point of greater opening, the oral channel is wide enough that even the high-velocity, voiceless airstream will not cause turbulence.[4] This marks the degree of stricture for resonants.

So, degree of stricture can be described as a continuous articulatory dimension of speech sounds which involves four different quantal acoustic regions: stop, fricative, approximant, and resonant. Since we can regard [degree of stricture] as a multivalent phonetic feature, much like the phonetic feature [place], one wonders if this phonetic dimension can be expressed in terms of a hierarchy of phonological features in a way analogous to the hierarchy used to describe place of articulation. I will approach this issue by considering the phonological features which may be used to describe degree of stricture, and the "taxonomic dependencies" which exist among them.

---

[2] This is true even for a voiced closure, provided the velar-pharyngeal port is also closed and a relatively short amount of time passes. With the passage of time during the production of a voiced stop air pressure above and below the glottis is equalized, and consequently airflow through the glottis ceases.

[3] Stevens (1972) defined approximants as sounds for which "vocal-cord vibration can be maintained only through an adjustment in the mode of vibrating." This is related to Catford's definition in that the oral constriction in approximants is enough to impede airflow and thus, to create increased intra-oral air pressure which hinders voicing.

[4] That is, the voiceless airstream will not result in turbulence at the narrowest point of the oral tract. Voiceless sounds *are* characterized by turbulence at the glottis, however.

Features for Degree of Stricture. The features which describe the different degrees of stricture are: [consonantal], [sonorant], and [continuant]. [5] Among these features there seem to be some taxonomic dependency relations similar to those noted above among the (SPE) features for place of articulation.

[Consonantal] is defined by Halle and Clements (1983) as "produced with a sustained vocal tract constriction at least equal to that required in the production of fricatives", while [sonorant] is defined as "produced with a vocal tract sufficiently open that the air pressure inside and outside the mouth is approximately equal". These features mark off overlapping regions along the degrees of sticture dimension in much the same way that [anterior] and [coronal] do along the place dimension. [Consonantal] separates vowels from the other types of segments, while [sonorant] distinguishes among the consonants, and is essentially redundant for vowels. Note also that only one value of the feature [sonorant] is interpretable for [-consonantal] sounds. There is no class of sounds which can be defined as [-cons][-son]. This is further indication that [sonorant] need be specified only for [+consonantal] sounds.

Among the consonants [continuant] serves two purposes. First, [continuant] separates stops from fricatives. Second, it is used to distinguish two classes of sonorant consonants: the nasals and the rhotic sounds. Since this second distinction is also accomplished by the feature [nasal], the primary function of [continuant] is to classifiy stops separately from fricatives. This means that it needs to be specified only for those sounds which are [+consonantal] and [-sonorant]. Given these considerations the four degrees of stricture mentioned above can be described by the feature specifications in (2).

$$\begin{array}{ll} \textit{resonant} & [-cons] \\ \textit{approximant} & [+cons], [+son] \\ \textit{fricative} & [+cons], [-son], [+cont] \\ \textit{stop} & [+cons], [-son], [-cont] \end{array} \tag{2}$$

The implicit dependencies among the features for degree of stricture are shown in (3). This type of representation is different from the hierarchical representation used to describe place of articulation in several ways. First, there is no class node. If we were to adopt the type of representation of manner in Clements (1985) or Ladefoged and Halle (1988) the structure in (3) might appear under a manner (or stricture) node, but as it stands there is no internal class node in the structure. Second, all of the features in the structure are binary. There are no privative features comparable to the articulator features in Sagey (1986b) or Ladefoged and Halle (1988).

---

[5] I am (at this point) using these features as they are defined in Halle and Clements (1983).

$$\begin{bmatrix} consonantal \\ + \\ sonorant \\ - \\ continuant \end{bmatrix} \qquad (3)$$

The structure in (3) can be interpreted as defining possible feature bundles by defining allowable paths through the features. The procedure for specifying a feature bundle can be described by the sequence of operations illustrated in (4). In this way [sonorant] only receives a value if the segment is also [+consonantal], and [continuant] only receives a value if the segment is [-sonorant].

$$\begin{array}{lll} specify(consonantal) & & \\ if(consonantal = plus) & then & specify(sonorant) \\ if(sonorant = minus) & then & specify(continuant) \end{array} \qquad (4)$$

Glides. Since glides are phonetically approximants, I have cheated in my use of the feature [consonantal] by grouping all approximants (including glides) as [+cons][+son].[6] Although it isn't phonetically necessary to define two types of approximants, there may be phonological differences between what we could call consonantal approximants and vocalic approximants. This could be captured in the structure in (3) by adding the feature [syllabic] as definable for [-consonantal] segments (i.e. *if (consonantal=minus) then specify(syllabic)*). This would produce the two types of approximants; [+cons][+son] and [-cons][-syll] (as a means of representing the difference between [j] and [ι], for example). This may, or may not, be a desirable result. If diphthongs should be considered single segments (with a branching specification for [syllabic]) then this may be a welcome addition to the descriptive apparatus. One wonders, though, if the functional distinction between consonantal and vocalic approximants might be better left to rules of prosodic structure.

Now, having considered the featural representation of degree of stricture, we must address the question of where in the segmental hierarchy these features should appear.

---

[6] I am not alone in this redefinition. Hyman (1985) argues that [consonantal] should be identified with the old distinction between "vocoids" and "contoids" (see Pike, 1943). He suggests that "glides are [+cons] *on the surface*" (p. 77, italics his), but that they may be derived from [-cons] (i.e. vocoid) segments.

Position in the Segmental Hierarchy

As mentioned above, there have been a number of suggestions concerning the location of manner features in the hierarchy of segmental features. It is possible to argue from a phonetic point of view that the features for degree of stricture (3) are fundamentally different from the features which refer to particular locations within the vocal tract. In the following, I discuss the phonetic uniqueness of the degree of stricture features and suggest a location for them in the segmental feature hierarchy.

Of the two major phonetic dimensions of speech sounds, place of articulation and manner of articulation, the place dimension can be described without reference to time, while the manners of articulation (and here I am referring to degrees of stricture in particular) must include some element of the passage of time.

So, for instance, a fricative, to be a fricative, must have the articulators in a configuration in which turbulence is generated for some minimal amount of time.[7] This can be compared to sounds classified as [coronal]. No matter how briefly the tongue is positioned on or near the alveolar ridge if the position is reached the feature applies. As an example of the necessary reference to time in the case of the feature [continuant], consider the frication which is produced when a stop closure is released. In going from [-continuant] to some more open variety of segment the articulators pass through a fricative configuration. This can be seen in spectrograms as a frication after the release of stop closure which has the spectral qualities of a fricative produced at that place of articulation (especially for aspirated stops). The implicit reference to time in the features for degree of stricture captures the fact that stop release frication does not "count" as a fricative. No other dimension separates release frication from fricative consonants. They are articulatorily and acoustically comparable to each other.[8]

The same type of argument can also be constructed for the feature [sonorant]. For instance, during the transition between stop and vowel, the articulators invariably pass through a configuration appropriate for a homorganic approximant, but this configuration does not count as an approximant segment unless the configuration for an approximant exists for some minimal duration.[9] Likewise, [consonantal] includes a reference to the time domain.

---

[7] Because "degree of stricture" has quantal regions, it is not necessary that the articulators be *stationary* in order for the acoustic output to be fairly constant. So, when I say the articulators must be "in in a configuration in which turbulence is generated for some minimal amount of time" I am not suggesting that the articulators must be stationary for some minimal amount of time.

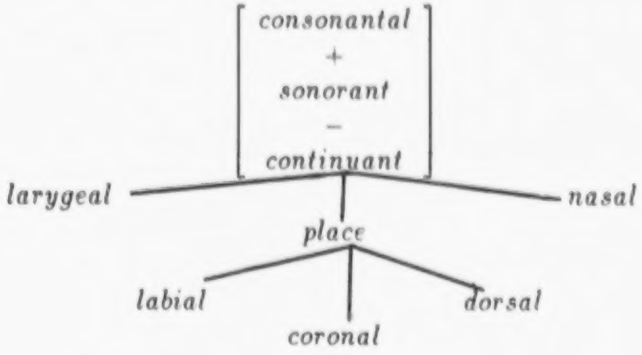
[8] There can be temporal variation among segments which have the same degree of closure. For instance, the primary difference between flaps and full stops is the duration of the closure. Ladefoged (1971) includes a phonetic parameter he calls "rate" which is used to distinguish stops, flaps, taps and trills.

[9] At first blush, the reduction of phonemes in fast speech seems to provide a counter-argument against these minimal duration arguments. In such cases we probably need to distinguish between underlying and phonetic segments. A segment may be underlyingly present in a lexical item even though not physically (and hence, phonetically) present at all in fast speech.

This is mentioned explicitly in Halle and Clements' (1983) definition of the feature [consonantal]: "produced with a sustained vocal tract constriction." So, there is a natural division of features into those which are time-based and those which are not.

These time-based features define the type of gesture made by the articulators, while the other features define the locations in the vocal tract of those gestures. This distinction is reminiscent of the distinction which McCarthy (1981) made between melody-bearing elements and melodic elements. In the proposal offered here, the melody-bearing elments are the gesture types and the melodic elements are the vocal tract locations of those gestures. In support of this distinction, consider the success which phonologists have had in the construction of a geometry of melody features as opposed to the degree of stricture features. The phonological criteria proposed by Clements (1985) have succeeded in identifying the melodic features because these properties of segments may be smeared across time; they participate in assimilations. Or, conversely, their range of application in time may be limited; they may delink. The features which define gesture types have been harder to pin down because they are time-bound and consequently don't fit the phonological criteria.

So, I am arguing for a distinction among features between those which are time-bound and those which may spread or delink. Following Mester (1986) and McCarthy (1988), this distinction can be represented in a feature hierarchy by placing the time-bound features in the root node of the feature hierarchy (see (5)). Mester's (1986) description of this feature bundle as the segment "core" is especially appropriate. The core describes the gesture type and the structure linked to it describes the location within the vocal tract of those gestures. Thus, the proposal that I am offering is that the structure in (3) and, more particularly, the feature bundles which it defines, is the root node of the segmental feature hierarchy.



(5)

Time-bound Features and the Timing Tier. If the features for degree of stricture are time-bound, what is the relation between them and the timing tier? In answer to this question, I think it is useful to consider two types of timing: segmental timing and prosodic timing. Segmental timing is determined by properties internal to the segment. In particular, I propose that segmental timing (to a rough approximation) is determined by the gesture type.[10] Prosodic timing, on the other hand, is determined by the position of the segment in a prosodic structure. So, location in a syllable or foot, and the structure of those prosodic elements, plays a role in the durational properties of the segment. In view of this distinction, we can then assert that the segment core (3) *is* the segmental timing tier, and stipulate that time-bound segments are restricted to the segmental timing tier.

A number of types of phonological rules can be expressed as linking or delinking to this segmental timing tier. These rules include vowel harmony, nasal spreading, voicing assimilation, and the palatalization or labialization of consonants. Other rules require that there be a prosodic structure within which segments reside. These include allophonic variation due to location within a syllable as found with aspiration for [-voi] stops in English and velarized [l] in English. From the prosodic template we may also derive the notion of a segmental "slot" which is characterized as a location within a prosodic structure, rather than by segmental features. This construct is necessary for the statement of processes of compensatory lengthening, for example.

Nasal and Lateral. The manner features [nasal] and [lateral] do not find a place among the degrees of stricture features in (3). The nasal consonants can be described as doubly articulated consonants, involving both an oral closure and a velo-pharyngeal approximation. Likewise, a phonetic description of laterals must mention both that the segments are produced laterally and that they have a particular degree of stricture (approximant or fricative). The articulatory fact which motivates these descriptions is that for both nasals and laterals there are two types of stricture at two places of articulation. For nasals in particular the two places of articulation are fully independent. The oral cavity may be in the configuration of a vowel, or consonant (the primary place of articulation) while the velar-pharyngeal port is open or closed (the secondary place of articulation).[11] Laterals may also be viewed as a combination of two articulatory gestures. First, a closure along the midline of the vocal tract, and second, an approximate or fricative along one or both sides to the vocal tract.

If we thus consider nasals and laterals to be doubly articulated segments, then whatever mechanism which is adopted for the representation of multiple articulation can also be used to represent the degree of stricture of the secondary articulation of nasals and laterals. For instance, we could use Sagey's (1986b) conception of major and minor articulators and manner pointers to specify the degree of stricture for the major articulator. [12] So, [m] can be represented as having [labial] as its major articulator and [nasal] as its minor articulator. If, following Sagey (1986b) degree of stricture is explicitly stated for the major articulator and degree of stricture for the minor articulator is determined by language specific rule, we can capture the articulatory fact that nasal consonants have two degrees of stricture (i.e. they are both stops and sonorants).

---

[10] I am not, here, attempting to account for details of articulator movement trajectories. My claim is simply that differences in overall segment duration are related to stricture type.

[11] The two limitations on configuration of the oral tract during nasals are: (1) the oral tract may not be closed below the velar-pharyngeal port (as in a pharygeal stop) because airflow then can't go through the nose, and (2) the air pressure required to produce a fricative cannot be produced (at normal levels of subglottal air pressure) when the nose is also open. In both of these cases, it is articulatorily possible to produce the offending combination of gestures, it just turns out that the combination is ineffective as a means of producing identifiable speech sounds.

In describing laterals in this way, we end up in the odd position of asserting that [coronal] is both the major and minor articulator; [coronal] stop along the midline of the vocal tract and [coronal] approximant or fricative along the side of the tongue. Notice, however, that this description of laterals provides a simple way to represent the distinction between lateral approximants and fricatives.

Place Information in the Segment Core. Stricture features are time-bound, are they also location independent? Currently, [cons], [son] and [cont] are defined for the vocal tract and not for the larynx. If we assert that the features in the segment core are not specified for place of articulation, we then increase their domain of application to the larynx as well. This may be useful for the statement of debuccalization. In this process, if the vocal tract features are delinked from a segment the segment retains its specification for stricture type, but the stricture is realized at the larynx. For instance, [k] – > [ʔ] and [x]– > [h]. This is a straight-forward process if the stricture types are independent of location in the vocal tract.[13]

The phonological motivation for considering [ʔ] and [h] as [+cons][-son] was to group them in a natural class with the vowels. If we stipulate that the larynx is a secondary articulator, the vocal tract configuration is left as the primary articulator and thus, the degree of stricture in the vocal tract is the same for both vowels and glottal consonants.[14]

---

[12] I am not making a commitment to any particular representation of primary/secondary articulation. It might be possible to use Selkirk's (1988) suggestion that secondary articulation be represented in terms of a dependency relation between articulators. In this approach, the doubly articulated stop [k͡p] would then be represented with [labial] under the [dorsal] node if the segment is a labial-velar, or with [dorsal] under the [labial] node if the segment is a velar-labial. Then, if we consider [nasal] to be a secondary place of articulation we can stipulate that [nasal] is a "perpetual dependent" (something like a graduate student). One problem with this is that, as Selkirk pointed out, the notion tier in this proposal is defined in terms of dependency relations. This is a problem for a perpetually dependent feature [nasal] because in order to describe nasal spreading we need to refer to a nasal tier. If nasal is dependent on a variety of different place nodes, then it is not clear that there *is* a nasal tier. Maddieson and Ladefoged (1988) propose to represent secondary articulation by a separate node under the place node which they call the [secondary articulation] node. To treat [nasal] as a secondary articulation in this approach, [nasal] would reside under the [secondary articulation] node. However, their assertion that only one secondary articulation node is needed because only two places of articulation need to be specified in the description of distinctive speech sounds would not hold for doubly articulated nasals like [m͡ŋ].

[13] Delinking place creates an interesting situation. We can't really say that the vocal tract isn't involved in the production of [ʔ] and [h], yet the shape of the vocal tract is not distinctive in these sounds. This underlies the phonetic description of [h] before vowels as a voiceless version of the following vowel ([VV]).

[14] This is not an unmotivated stipulation. Like the velar-pharyngeal port ([nasal]), the larynx is articula-

Conclusion

The principal suggestions of this paper are: (1) Manner of articulation can be described as various combinations of degree of stricture and multiple articulation, (2) degree of stricture can be described by a hierarchical organization of the three features [consonantal], [sonorant] and [continuant], and (3) the specification of degree of stricture (for the primary articulator) is also a definition of "segment type" and best fits in a segmental feature hierarchy as the root node of the hierarchy.

Whether this view of manner of articulation can be useful in the description of phonological processes has been only briefly discussed. The examples discussed seem to indicate some degree of success, but there are no doubt other cases with which this view will not be so successful. Perhaps, for low-level, phonetically-motived phonological processes a representation such as the one I have proposed here will be most suitable, while in the description of higher-level phonological rules, which may refer to word classes and/or grammatical conditioning, a different type of (abstract?) representation will be more useful. If this is the case, then we are led to a formulation of the phonetics-phonology interface which includes the transformation of a (phonetically sparse) phonological representation into a representation such as that suggested here in which phonetic detail is more fully represented.

---

torily independent of the oral tract. There is at least one phonetic limitation on the degrees of stricture of the larynx: the larynx may not open widely enough to avoid the production of turbulance (at the rates of airflow found in speech). Thus, only closure, frication, and trill (voicing) are possible degrees of stricture for the larynx.

References

Anderson, S. R. (1976). On the description of multiply-articulated consonants. *Journal of Phonetics*, **4**, 17–27.

Catford, J. C. (1977). *Fundamental problems in phonetics.* Bloomington, IN: Indiana University Press.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English.* New York, NY: Harper and Row.

Clements, G. N. (1985). The geometry of phonological features. *Phonology Yearbook*, **2**, 225–252.

Clements, G. N. (1987). Phonological feature representation and the description of intrusive stops. *CLS Parasession*, **23**, 29–50.

Halle, M. & Clements, G. N. (1983). *Problem book in phonology.* Cambridge, MA: MIT Press.

Hyman, L. M. (1985). *A theory of phonological weight.* Dordrecht: Foris.

Ladefoged, P. (1971). *Preliminaries to linguistic phonetics.* Chicago, IL: University of Chicago Press.

Ladefoged, P. (1975, 1982). *A course in phonetics.* New York, NY: Harcourt Brace Jovanovich.

Ladefoged, P. & Halle, M. (1988). Some major features of the International Phonetic Alphabet. *Language*, **64**, 577–582.

Maddieson, I. & Ladefoged, P. (1988). *Multiply articulated segments and the feature hierarchy.* Paper presented at the 1988 LSA meeting.

McCarthy, J. J. (1988). Feature geometry and dependency: A review. *Phonetica*, **43**.

Mester, R. A. (1986). *Studies in tier structure.* Unpublished doctoral thesis, University of Massachusetts, Amherst.

Pike, K. L. (1943). *Phonetics.* Ann Arbor, MI: The University of Michigan Press.

Sagey, E. C. (1986a). On the representation of complex segments and their formation in Kinyarwanda. In Wetzels, L. & Sezer, E. (Eds.), *Studies in compensatory lengthening*. Dordrecht: Foris.

Sagey, E. C. (1986b). *The representation of features and relations in non-linear phonology.* Unpublished doctoral thesis, MIT.

Selkirk, E. O. (1988). *Dependency, adjacency, and secondary articulation.* Paper presented at the 1988 LSA meeting.

Stevens, K. N. (1972). The quantal nature of speech: Evidence from articulatory-acoustic data. In David, E. E. & Denes, P. B. (Eds.), *Human communication: A unified view.* New York, NY: McGraw-Hill.

Determining the Locus of Talker Variability Effects on the Recall of Spoken Word Lists: Evidence from a Presentation Rate Manipulation[1]

Stephen D. Goldinger, John S. Logan, and David B. Pisoni

Speech Research Laboratory
Department of Psychology
Indiana University
Bloomington, IN 47405

---

[1] This research was supported by NIH Research Grant *NS-12179-12* to Indiana University in Bloomington, IN. We thank M. Kimberley Spence for her help in running subjects and scoring data.

Abstract

In a recent study, Martin, Mullennix, Pisoni and Summers (in press) reported that subjects' accuracy in recalling lists of spoken words was better in early list positions when the words were spoken by a single talker than when the words were spoken by multiple talkers. Martin et al. discuss two possible explanations for their findings. One explanation suggested that the perceptual costs associated with talker variability usurp processing time and resources necessary for rehearsal. The second explanation suggests that, in addition to perceptual costs, talker variability interferes with the speed and efficiency of the rehearsal processes themselves. The present study was conducted to determine whether the locus of talker variability effects is confined only to the early perceptual encoding of items, or if talker variability also affects the efficiency of rehearsal processes as well. Accuracy of serial-ordered recall was examined for lists of words spoken by a single talker and by multiple talkers that contained either "easy" words or "hard" words. Rate of presentation of the items was manipulated over a wide range to determine if either of these variables affect rehearsal processes as well as perceptual encoding. Across all conditions, recall of list items improved as presentation rates decreased. A strong interaction was obtained between talker variability and rate of presentation. Recall of multiple-talker lists was affected much more than single-talker lists, suggesting that talker variability affects not only perceptual encoding, but also some aspect of the rehearsal process. No such interaction was obtained between word confusability and rate of presentation. The data provide support for a proposal offered by Martin et al. that talker variability affects the accuracy of recall of spoken material not only by increasing the processing demands for early perceptual encoding of the input but also by reducing the efficiency of rehearsal processes used to transfer items into long-term memory.

Determining the Locus of Talker Variability Effects on the Recall of Spoken Word Lists: Evidence from a Presentation Rate Manipulation

The perception of spoken language is a feat that requires the listener to extract stable linguistic content from a physical signal that is notoriously unstable. The acoustic realization of speech is simultaneously modulated by numerous variable phonetic, prosodic and semantic characteristics inherent to the particular message. The speech signal is further modulated by varying source characteristics. Indeed, if the perceptual system is required to extract canonical units of meaning from speech signals, various idiosyncracies related to different talkers must be normalized during the earliest stages of processing the sensory input. Such idiosyncracies include differing individual dialects, vocal tract shapes, and speaking rates (Mullennix, Pisoni, & Martin, in press).

Casual observation of language performance leaves the impression that listeners can attend to several different voices in succession with virtually no perceptual costs or consequences. However, several studies suggest that speech perception and word recognition are impaired by talker variability.[1] For example, Summerfield and Haggard (1973) showed that talker variability impairs vowel perception. More recently, Mullennix et al. (in press) have shown that talker variability is harmful to subjects' word recognition performance. Mullennix et al. found that when stimulus items were presented in voices that changed from trial to trial, recognition was less accurate in a perceptual identification task and was slower in a naming task. These findings led Mullennix et al. to conclude that some resource-demanding mechanism is used by listeners to compensate for variations in speakers' voices.

Other lines of evidence suggest that talker variability affects not only speech perception, but memory processes as well. Recent experiments conducted by Martin et al. (in press), Mullennix et al. (in press) and by Logan and Pisoni (1987) have shown that word lists produced by multiple talkers are more difficult to recall than word lists produced by a single talker. In a series of experiments, Martin et al. found that: 1) Serial-ordered recall of spoken word lists was worse for multiple-talker lists than for single-talker lists, but only for items from early list positions, 2) recall of visually presented digits presented *before* the presentation of the spoken lists was worse if the subsequent lists were multiple-talker lists than if they were single-talker lists, and 3) these differences in primacy recall were unaffected by a post-perceptual distractor task (following Peterson & Peterson, 1959). From these converging lines of evidence, Martin et al. suggested that word lists produced by multiple talkers may require greater processing resources for rehearsal in working memory than lists produced by a single talker.

---

[1] Throughout this paper, the terms "talker variability" and "talker condition" are used. Although the term could mean several kinds of variability, we are using it here only to denote situations in which spoken items are produced by *different* talkers from trial to trial. We do not intend the term do denote variability of vocal quality *within* a talker.

The explanation offered by Martin et al. is closely related to the obligatory and attention-demanding nature of voice information in speech perception. Such a position is supported by recent findings by Mullennix and Pisoni (1987), who employed the Garner (1974) speeded classification paradigm. They found that information about a talker's voice is processed in parallel with the phonetic content of words in an integral fashion. That is, subjects were unable to selectively ignore voice information, even when attending to voice information was detrimental to performance in the primary classification task. Apparently, changes in the voices of talkers from trial to trial in speech perception tasks requires continual reallocation of selective attention. Similar findings have been reported by Geiselman and Bellezza (1976), who found that voice information for spoken material is retained, even when subjects received no specific instructions to attend to voice characteristics. Martin et al. have suggested that when listeners are required to memorize lists of words spoken by multiple talkers, the variability from different talkers demands additional processing resources that are needed for the efficient rehearsal and transfer of list items to long-term memory. As a consequence, recall of early list items is impaired.

While the processing capacity-based explanations offered by Martin et al. are consistent with their data, the authors note that there are actually two possible ways that talker variability could affect rehearsal processes. The first possibility (following Mullennix et al., in press) is that the locus of the effects is confined to early perceptual encoding. That is, the extra time and resources required to "normalize" each token in a multiple-talker list is time taken away from higher processing systems, reducing available rehearsal times for each token and therefore attenuating early list performance. The other explanation is that talker variability effects may also influence the rehearsal processes themselves. That is, in addition to the early perceptual costs, it is possible that the variability in voice information contained in each list makes the items more difficult to rehearse and encode into long-term memory.

The present study was conducted to evaluate these alternate explanations more closely. In particular, we wanted to determine whether talker variability affects only perceptual encoding or perceptual encoding *and* rehearsal processes. To study this, we manipulated two additional variables: stimulus confusability and rate of presentation. The stimulus confusability dimension was selected for its known influence on perception. Each word list contained ten words, half of which were "easy" words and half of which were "hard" words. The confusability variable used here is a combined metric of two measures known to influence word recognition, word frequency and lexical density. The first measure is based on the frequency count of Kučera and Francis (1967). The second measure is based on analyses of *similarity neighborhoods* (Luce, 1986). A similarity neighborhood is defined as a collection of words that sound similiar to a given word. One characteristic of similarity neighborhoods that has been examined is the number of neighbors any given word has; some words have many similar-sounding neighbors whereas other words have fewer neighbors. For spoken words, Luce (1986) has shown that word recognition is slower and less accurate for words selected from dense neighborhoods than for words selected from sparse neighborhoods. In the present study, "easy" words were defined as high frequency words selected from sparse neighborhoods, whereas "hard" words were defined as low frequency words selected from dense neighborhoods.[2]

The talker manipulation was the same in the present study as the in the Martin et al. study: Talker variability was manipulated as a between-subjects variable; some subjects heard all single-talker lists whereas other groups heard all multiple-talker lists. The confusability manipulation was a within-subjects variable; half of the lists for each group were comprised of "easy" words and half were comprised of "hard" words.

Casual observation suggests that there are qualitative differences that distinguish the stimulus dimensions of talker variability and word confusability. While it has been demonstrated that both talker variability and word confusability influence early perception, it is not clear whether both dimensions will affect rehearsal processes directly. The findings of Mullennix and Pisoni (1987) and Geiselman and Bellezza (1976) suggest that qualities inherent in voice information are not only attention-demanding, but are perceptually salient and potentially useful as well. Voice information conveys important information about a speaker's gender and emotional state (Geiselman & Bellezza, 1976). It is not clear that information regarding word confusability shares these characteristics with voice information. Indeed, it is not even clear that subjects have any reliable intuitions or categories to distinguish words on an abstract scale such as "confusability." Consequently, the perceptual and elaborative resources that are apparently dedicated to processing voice information may not be applicable to processing abstract information regarding word frequency or confusability. If this distinction applies, we would expect that talker variability would interact with post-perceptual rehearsal processes but that word confusability would not.

In order to determine whether talker variability and word confusability both impact upon memory processes in the same way, we examined the accuracy of recall for word lists varying along both stimulus dimensions across five levels of a third experimental manipulation. Specifically, because we were interested in a manipulation that would primarily affect rehearsal processes, we varied the rate of presentation of words in each list (Murdock, 1962; Rundus, 1971). Words were presented to subjects at one of five rates, with inter-word intervals of either 250, 500, 1000, 2000, or 4000 msec.

If talker variability affects rehearsal efficiency as well as perceptual encoding, there should be a strong interaction of talker condition and rate of presentation. Changes in rate should affect recall of words from multiple-talker lists more than from single talker lists. Furthermore, if word confusability affects perceptual encoding but leaves rehearsal processes relatively unaffected, there should be no interaction of confusability and rate of presentation. Changes in rate should affect recall of words from easy and hard lists equivalently.

---

[2] Throughout the remainder of this paper, for the ease of composition and comprehension, the following terminology is employed: The stimulus dimension relating to single versus multiple talkers will be referred to as the "talker" variable or manipulation. Similarly, the stimulus dimension relating to "easy" versus "hard" words will be referred to as the "confusability" variable or manipulation.

Method

Subjects. One hundred and sixty students enrolled in an introductory psychology course at Indiana University served as subjects. Subjects received course credit for their participation. All subjects were native speakers of English and reported no history of a speech or hearing disorder at the time of testing.

Stimuli. The stimuli were obtained from a large digital database of spoken monosyllabic words recorded by several different talkers. This was the same source used by Martin et al. (in press). The original words came from the vocabulary used in the Modified Rhyme Test (House, Williams, Hecker, & Kryter, 1965). In the present experiment, only a subset of the original 300 words were used. The words selected for the present experiment were those that satisfied several necessary constraints: First, the words were ranked according to their frequency of occurrence according to the Kučera and Francis (1967) norms. Second, the words were ranked according to their neighborhood densities, as determined by a one-phoneme substitution, addition, and deletion metric (Luce, 1986). Third, words were also ranked according to their neighborhood frequencies, a measure of the average frequency of the words' neighbors. Using these three criteria, two sets of words were selected for use in the present experiment. One set, the "easy" words, consisted of high frequency words from low-density, low frequency neighborhoods. The other group of words, the "hard" words, consisted of low frequency words, from high-density, high frequency neighborhoods. A final criterion used in selection was subjective familiarity; all of the words chosen for use in the present experiment were rated as highly familiar by subjects in a previous experiment conducted by Nusbaum, Pisoni and Davis (1984). After the words were divided into "easy" and "hard" sets according to these four criteria, each set contained 50 items. These 100 words were then used to generate 10 lists of ten words each. Five of the lists contained "easy" words and five contained "hard" words.

Once the words had been selected, digitized files containing tokens of each word were obtained from the database. One set of tokens was chosen from utterances produced by a single male talker; these tokens were used for the single talker conditions of the experiment. Another set of tokens was selected from the database so that every word in each list was spoken by a different talker; these tokens were used for the multiple talker conditions. In the multiple talker conditions, the same ten talkers, five males and five females, were used for all ten lists of words. The talkers used in the present experiment were the same talkers used in the Martin et al. (in press) study. All of the stimuli were originally recorded on audio tape and were then digitized with a 12-bit analog-to-digital converter using a PDP 11/34 computer. The mean RMS amplitude of all stimulus tokens was equated using a signal processing package.

Procedure. Subjects were tested in groups of six or fewer in a quiet testing room used for speech perception experiments. Stimuli were presented over matched and calibrated TDH-39 headphones at 75 dB SPL. A PDP 11/34 computer was used to present the stimuli and to control the experimental procedure in real-time. The digitized stimuli were reproduced using a 12-bit digital-to-analog converter and were low-passed filtered at 4.8 kHz.

All subjects were tested under the same conditions. Subjects first heard a 500 msec 1000 Hz warning tone indicating that a list of words was about to be presented. Then, a list of ten words was presented at one of five rates: one word was presented either every 250, 500, 1000, 2000, or 4000 msec. The presentation rate selected was held constant for any given group of subjects for the entire experiment. After each list of words, another tone was presented, indicating the beginning of the recall period. Subjects had 60 seconds to recall all the words they could. The end of the recall period was indicated by the presentation of a third tone. Subjects were instructed to recall the words in the exact order of their presentation in the lists. Subjects wrote their responses in specially prepared answer booklets using pen or pencil.

Rate of presentation and talker condition were between-subjects variables; word confusability was a within-subjects variable. Thirty-two subjects were tested at each rate of presentation. Half of the subjects in each group were tested with single-talker lists and half were tested with multiple-talker lists. The same words were heard by all subjects; only the number of talkers and the presentation rates varied between subjects. The order of presentation of words within each list varied randomly from session to session. The lists themselves were presented in the same order in all conditions of the experiment; the presentation of lists for each group alternated between those lists containing "easy" and those containing "hard" words.

Results

Subjects' responses were scored as correct if the target word or some phonetically equivalent spelling of the target word was recalled in the same serial position as the items presented in the lists. The upper panel of Figure 1 shows the percentage of correctly recalled words as a function of serial position and talker condition, collapsed across presentation rate and word confusability. The lower panel shows the percentage of correctly recalled words as a function of serial position and word confusability, collapsed across presentation rate and talker.

Insert Figure 1 about here

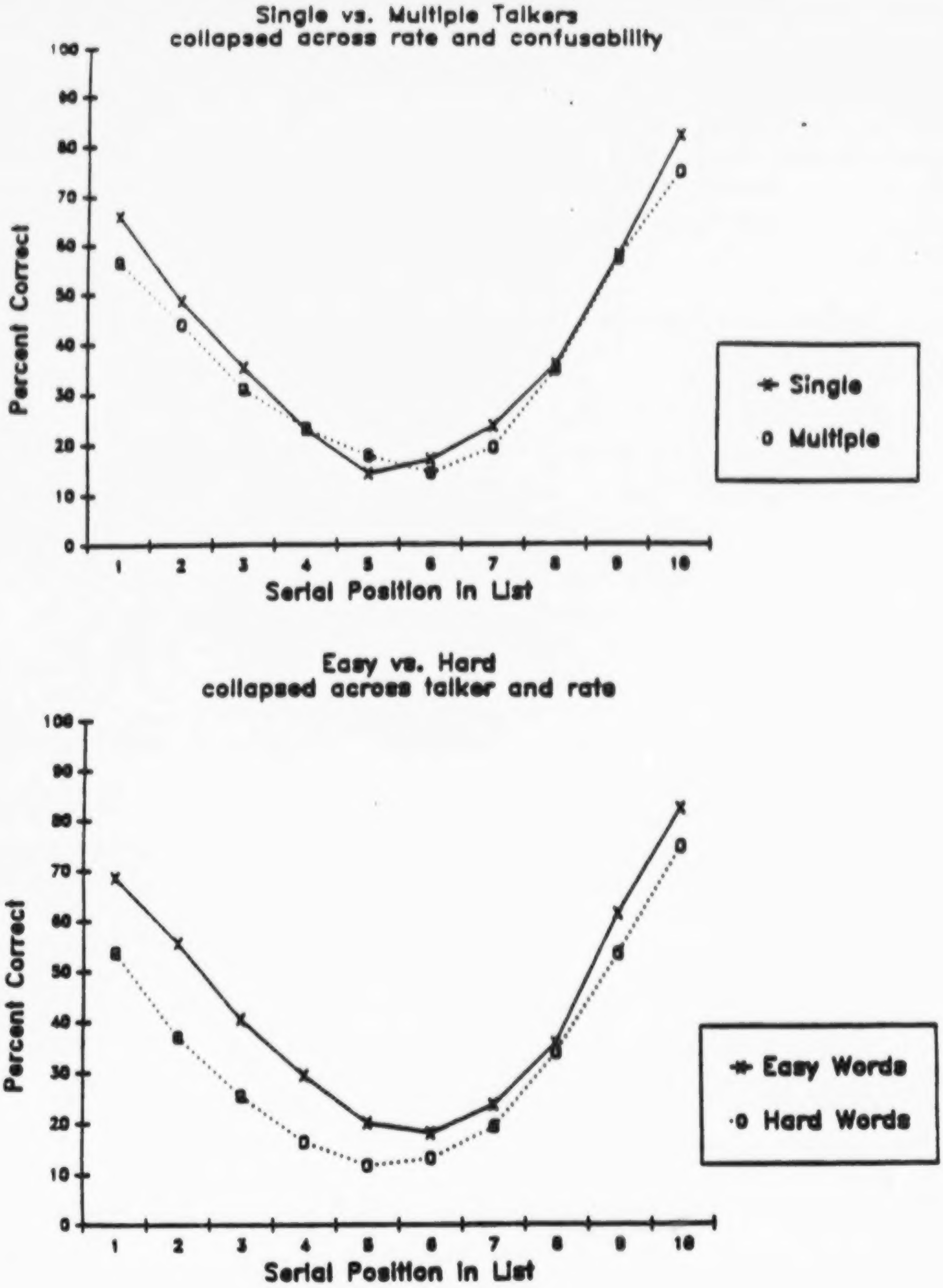


Figure 1. The upper panel shows mean percentages of correctly recalled words as a function of serial position and talker condition, collapsed across presentation rate and word confusability. The lower panel shows mean percentages of correctly recalled words as a function of serial position and word confusability, collapsed across presentation rate and talker.

A four-way analysis of variance (talker X word confusability X serial position X rate of presentation) was conducted on the percentages of correct responses. As expected, there was a significant main effect of talker $[F(1,150) = 3.98, p < .05]$. Words spoken by a single talker were recalled more accurately than words spoken by multiple talkers. In addition to the effect of talker, Figure 1 also shows a strong main effect of serial position $[F(9,1350) = 267.00, p < .0001]$, reflecting the usual U-shaped function obtained in recall tasks. A significant two-way interaction of talker and serial position was also obtained, $[F(9,1350) = 2.05, p < .05]$. The differences between the single-talker and multiple-talker recall functions tend to be larger at earlier list positions. Post-hoc Tukey's HSD analyses were performed on the percentages of correctly recalled words at each serial position. By these analyses, the recall functions for single and multiple-talker lists were significantly different only at serial positions 1 and 10. The main effect of talker obtained here was smaller than the effect reported by Martin et al. (in press). However, the main results of interest are obscured by averaging over the presentation rate manipulation, as will be discussed below.

A significant main effect of word confusability was obtained $[F(1,150) = 147.70, p < .0001]$. Recall of "easy" words was more accurate than recall of "hard" words at most serial positions of the lists. There was, however, a significant two-way interaction of word confusability and serial position $[F(9,1350) = 8.29, p < .0001]$, reflecting the larger differences between the recall functions for "easy" and "hard" words at early list positions. Post-hoc analyses showed that accuracy of recall for "easy" and "hard" words was significantly different at serial positions 1, 2, 3, 4, 5, 9, and 10.

Figure 2 shows data for the single and multiple talker conditions as a function of serial position and rate of presentation, collapsed across word confusability. The upper panel displays the recall functions for lists spoken by a single talker at five rates, ranging from Very Slow to Very Fast. The lower panel displays the recall functions for lists spoken by multiple talkers at the same five rates.

Insert Figure 2 about here

The ANOVA revealed a significant main effect for the rate of presentation $[F(4,150) = 22.56, p < .0001]$. Word recall improved as the rate of presentation became slower. More importantly, however, there was a significant three-way interaction of talker, rate of presentation, and serial position $[F(36,1350) = 2.23, p < .0001]$. This interaction reflects the tendency for the rate manipulation to affect recall of items from the primacy portions of multiple-talker lists *more* than single-talker lists. As the rate of presentation changed, recall of items from multiple talker lists was affected more than recall of items from single talker lists.

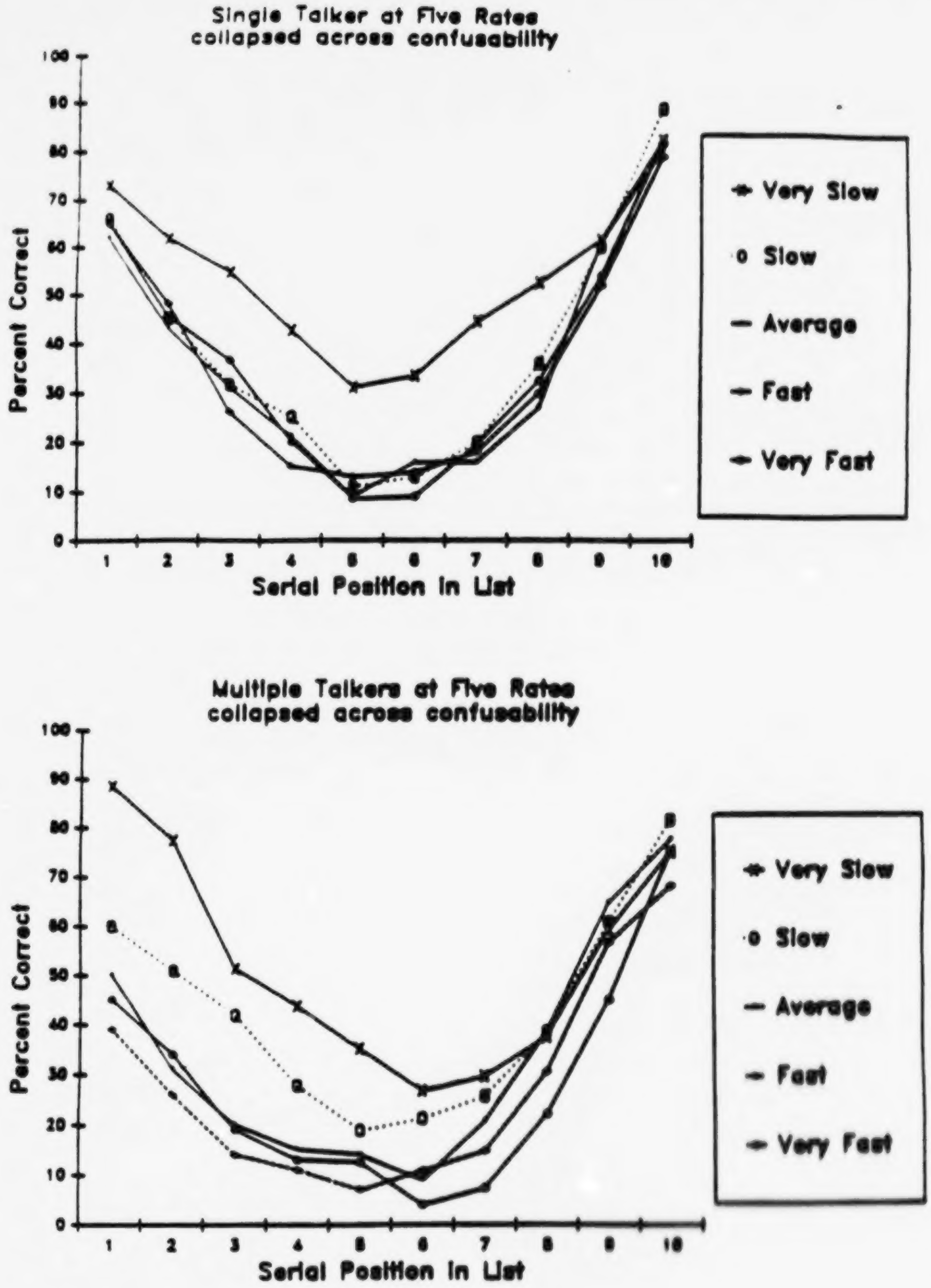


Figure 2. Mean percentages of correctly recalled words for both the single and multiple talker lists as a function of serial position and rate of presentation, collapsed across word confusability.

Figure 3 shows recall of both the "easy" and "hard" word lists as a function of serial position and rate of presentation, collapsed across talker. The upper panel displays the recall functions for lists of "easy" words spoken at five rates; the lower panel displays the recall functions for lists of "hard" words spoken at five rates.

Insert Figure 3 about here

Although it is clear that the recall functions for both "easy" and "hard" words are affected by the presentation rate manipulation, the critical three-way interaction of word confusability, rate of presentation, and serial position was not significant in this analysis [$F(36,1350) = 1.14$, $p = .266$]. Thus, unlike the finding obtained for the talker manipulation, changes in rate of presentation did not differentially affect the recall of "easy" and "hard" words.

The effects of presentation rate may be seen more clearly in Figures 4 and 5. Figure 4 shows the recall functions for lists of words spoken by single and multiple talkers, collapsed across word confusability, at each of the five presentation rates.

Insert Figure 4 about here

The interaction with rate of presentation can be seen more clearly in this figure. At the faster rates of presentation, the accuracy of recall for single-talker lists was better than recall for multiple-talkers lists, especially in the primacy portion of the curves. As the rate of presentation decreases, however, the differences between the two recall functions diminishes, and eventually *reverses* at the slowest rate. Indeed, at the slowest rate, recall for early list items from the multiple-talker lists is actually *better* than recall for the single-talker lists. Post-hoc analyses were conducted to compare the recall functions at all serial positions. These analyses showed that in all conditions, with the exception of the Slow condition, the differences obtained in the early list positions were statistically reliable. In the Slow condition, significant differences in recall were observed only at positions 3, 4, 5, and 6. Items from multiple-talker lists were recalled better than items from single-talker lists. This crossover effect is responsible for the three-way interaction observed between talker, rate of presentation, and serial position noted above.

Figure 5 shows the recall functions for lists of "easy" and "hard" words, collapsed across talker, at each rate of presentation.

Insert Figure 5 about here

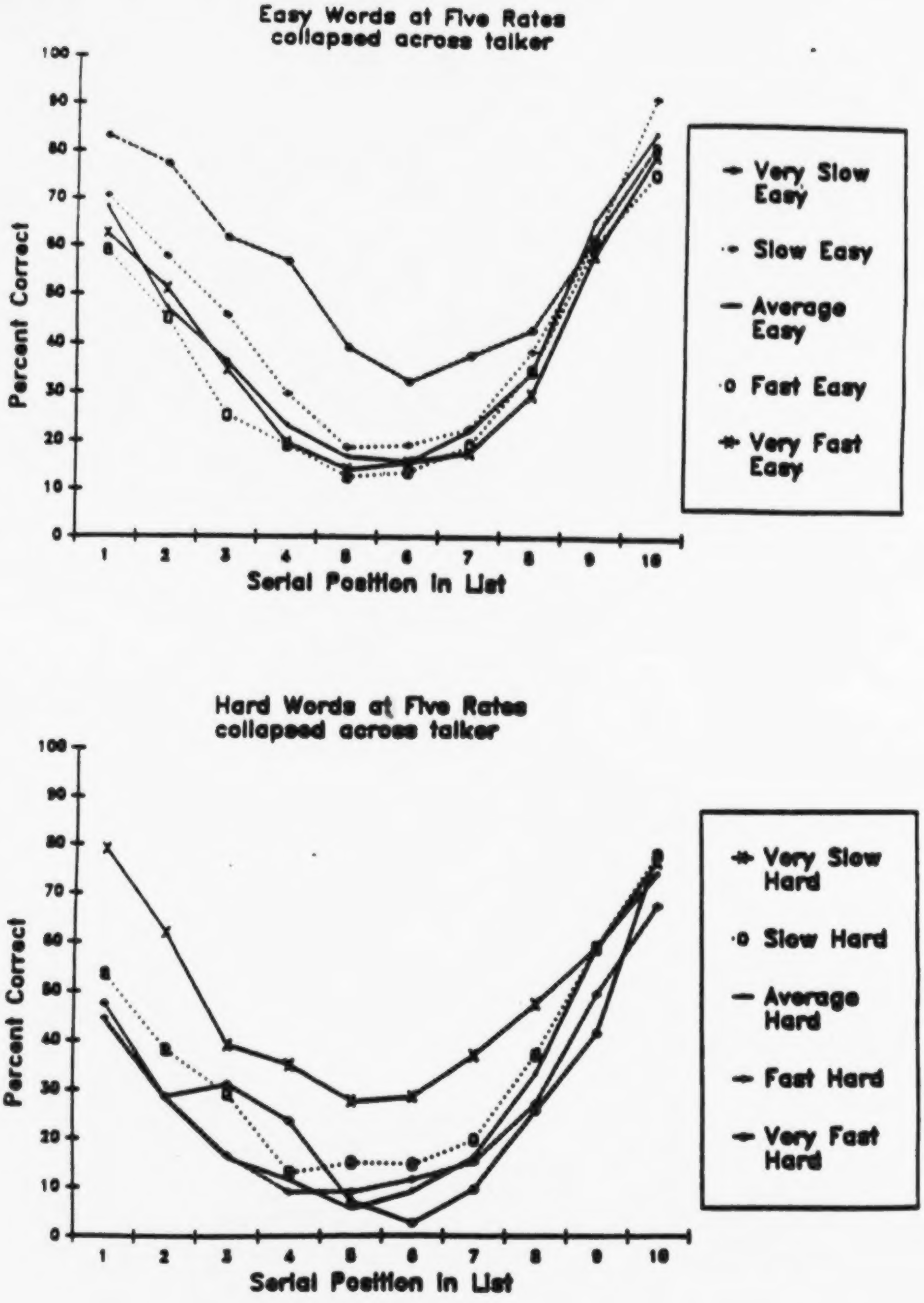


Figure 3. Mean percentages of correctly recalled words for both the "easy" and "hard" word lists as a function of serial position and rate of presentation, collapsed across talker.

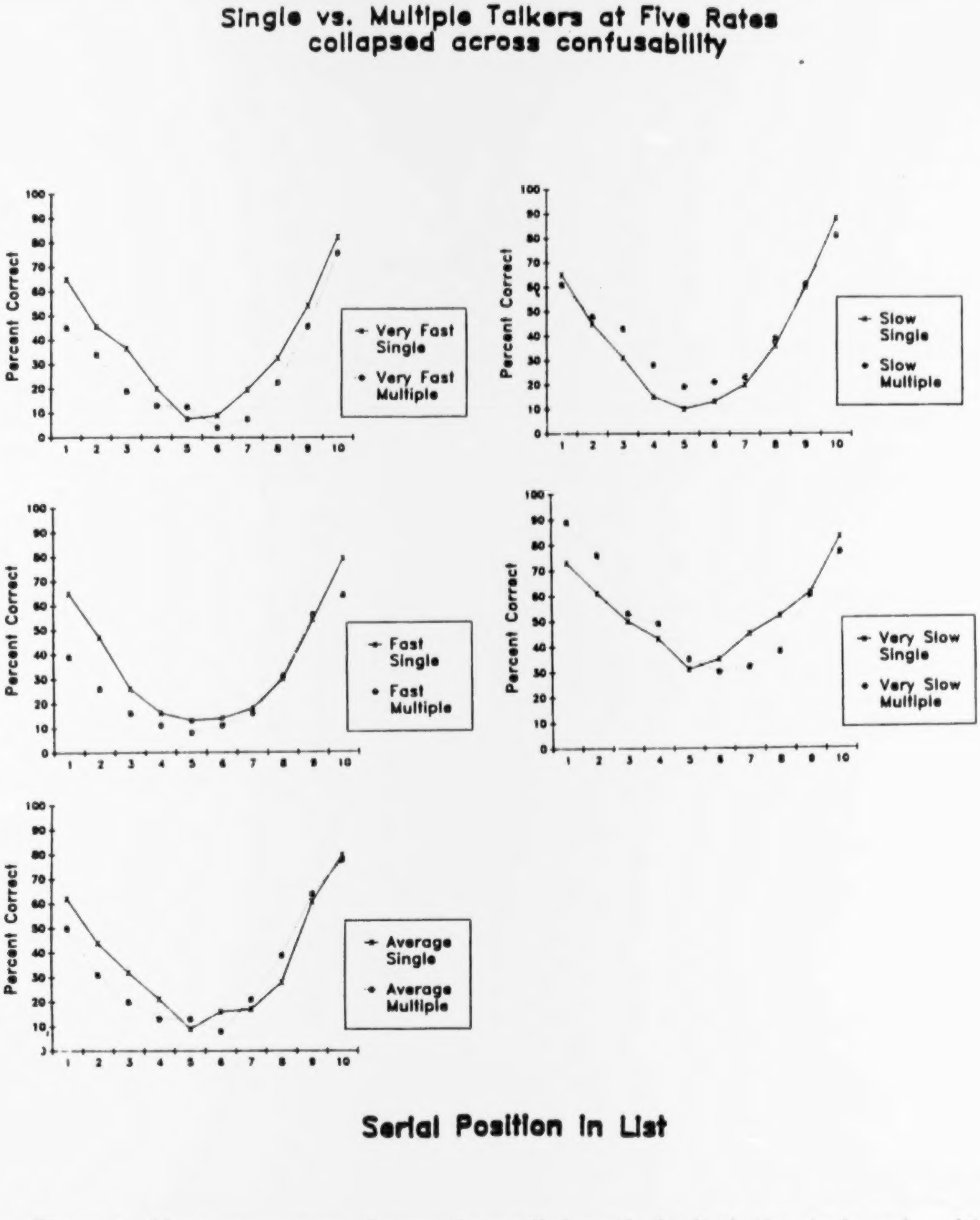


Figure 4. Mean percentages of correctly recalled words for both the single and multiple talker lists, collapsed across word confusability, at each rate of presentation.

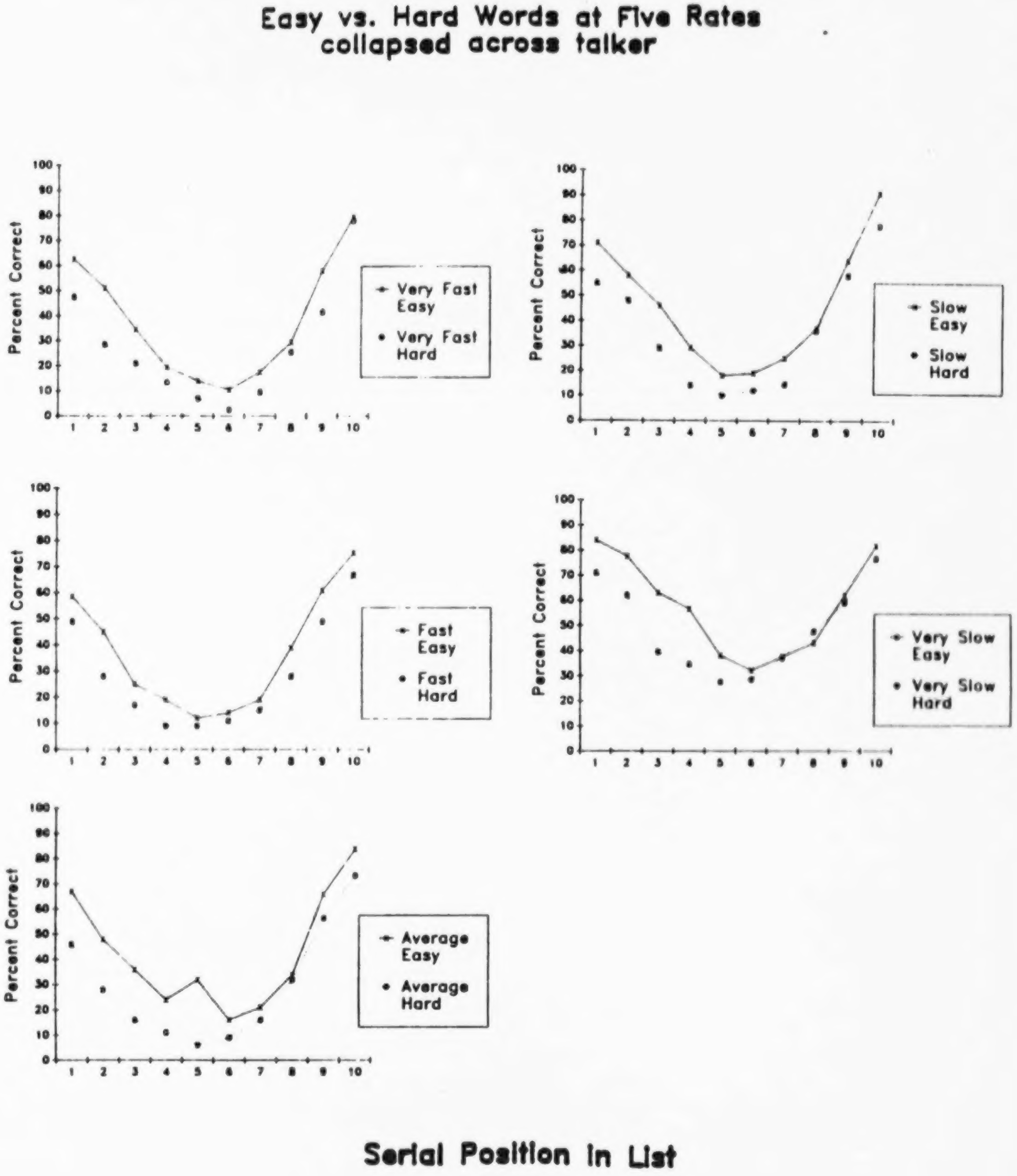


Figure 5. Mean percentages of correctly recalled words for both the "easy" and "hard" word lists, collapsed across talker, at each rate of presentation.

Examination of the recall functions for "easy" and "hard" words shows that changes in rate of presentation had comparable effects on both kinds of words. Although recall improves with slower rates, the improvement is comparable for "easy" and "hard" lists. These results are in marked contrast to the data shown in the previous figure. Post-hoc analyses revealed significant differences in recall performance for "easy" and "hard" words at most of the early list positions, and at several terminal positions as well. Accuracy of recall was consistently better for "easy" words than for "hard" words. With respect to the rate of presentation manipulation, Figure 5 clearly shows that while accuracy of recall varied with rate of presentation, the changes were equivalent for both "easy" and "hard" words.

Discussion

The present study was conducted to further examine the effects of talker variability on recall of lists of spoken words. Specifically, this study was conducted to test two alternate explanations suggested by Martin et al. (in press) for their finding that recall was better in the early list positions for lists spoken by a single talker than for lists spoken by multiple talkers. One possible explanation, an encoding account, is based on the perceptual consequences of talker variability. This view suggests that the initial delays that occur in word recognition when voice information changes from trial to trial simply cascade up the system and reduce the time and processing resources available for rehearsal of items. The second explanation is a rehearsal account. This view suggests that, in addition to the perceptual costs associated with talker variability, changes in the voice of the talker from item to item in a word list affects the speed and efficiency of the rehearsal processes required to transfer items from working memory into long-term memory.

To distinguish between these alternative explanations, we examined the effects of presentation rate on recall of lists spoken by single and multiple talkers, and we compared these effects to the effects of presentation rate on recall of lists of "easy" and "hard" words. Rate of presentation was manipulated because the literature has established that presentation rate affects recall of items in early list positions by reducing the available time for rehearsal and transfer of items to long-term memory (e.g., Murdock, 1962; Glanzer & Cunitz, 1966). If both talker variability and word confusability affect the accuracy of recall in the same ways, both variables should have been affected equivalently by the rate manipulation. In other words, if both variables affect rehearsal processes, recall for both multiple-talker lists *and* lists of "hard" words should have changed equivalently with respect to their appropriate counterparts as the rate of presentation changed. Such a pattern of results was *not* obtained in the present study; instead, we found that recall for multiple talker lists changed much more than recall for single-talker lists following rate changes, whereas recall for lists of "hard" words changed no more than recall for "easy" words. Indeed, we observed a surprising sensitivity of multiple-talker lists to the rate changes— at the slowest rate of one word every four seconds, recall of early list items for words spoken by multiple talkers was actually *better* than recall of items spoken by a single talker.

It should be noted that, although the evidence is not as impressive, there *is* a suggestion in the present data that the word confusability variable may affect rehearsal processes as well. Although the word confusability variable did not interact with the rate manipulation, the differences in recall between "easy" and "hard" lists were consistently larger at early list positions than in late list positions. Similar findings were reported by Sumby (1963), who found that recall for high frequency words was better than recall for low frequency words, especially in early list positions. Sumby suggested that it may be easier to rehearse more familiar words and encode them into long-term memory. Other explanations are available as well. For example, the hypothesis we have examined here regarding the perceptual costs of talker variability may be the appropriate explanation of the larger primacy effects of word confusability. Although the explanation now seems too simplistic to account for the effects of talker variability on recall, it is possible that the early perceptual delays caused by "hard" words indirectly affect rehearsal processes merely by reallocating processing resources. Another explanation of the larger differences between "easy" and "hard" words in early list positions may be based on retrieval of items from long-term memory. Because "hard" words in this study were low frequency words with many higher frequency neighbors, it is more likely that ambiguity and confusions during retrieval would cause more errors for "hard" words than for "easy" words. Whether the effects of word confusability on rehearsal are direct or indirect is not the issue here, however. The important point is that, while there is some suggestion that word confusability *can* affect rehearsal processes, there is now much stronger evidence that talker variability *does* affect rehearsal processes. Indeed, the observed advantage for recall of lists spoken by multiple talkers at the slowest rate suggests that voice information is available to subjects throughout the task and that this information can be used to encode words in long-term memory. It is likely that this information is used to elaborate and code items for permanent storage.

The present finding that voice information affects more than simply early perceptual encoding has precedents in the literature. As Mullennix and Pisoni (1987) and Martin et al. (in press) have suggested, voice information may be processed in an obligatory manner, in the sense of Fodor (1983). When subjects are listening to words spoken by many voices, information about the different voices demands attention and processing capacities. It is not clear, however, whether the majority of the processing efforts are dedicated to intentionally ignoring voice information, or if they are allocated to encoding voice information as useful cues for preserving order information. At slow presentation rates, the latter explanation seems more appropriate, given the advantage observed in this study for recall of multiple-talker lists. The issue remains unclear for the faster rates. Other researchers have reported similar findings revealing the obligatory and persistent nature of voice information as it combines with linguistic information in speech perception. For example, Cole, Coltheart, and Allard (1974), and Allard and Henderson (1975) have found that reaction times are faster for voice and name matches between target and probe items than for name matches alone.

Furthermore, Craik and Kirsner (1974) found that voice information remains in memory and affects word recognition for at least two minutes. Several studies conducted by Geiselman and his colleagues (Geiselman & Bellezza, 1976, 1977; Geiselman & Glenny, 1977; Geiselman & Crawley; 1983) have examined the incidental storage of speakers' voices during the processing of linguistic information, and have found that voice information is encoded into long-term memory even when subjects are not instructed to attend to voices.[3] Finally, in a somewhat whimsical study, Kosslyn and Matt (1977) found that subjects' knowledge of a writer's *speaking* rate can affect how quickly they read his or her prose.

Although the results of the present study strongly suggest that changes in voice information affect the rehearsal process, as well as the early perceptual encoding of list items, our claims could be fortified by more perceptual data. Because we have argued that both talker variability and word confusability affect perceptual encoding but only talker information strongly affects rehearsal, it is important to examine the relative perceptual costs of both manipulations. It may be that for the stimulus items used in the present study, the talker variability manipulation was far more damaging to perception than the word confusability manipulation. If this were the case, the implications of present data would be less clear. This is not to say the present findings would be uninteresting; the crossover effect found for single- vs. multiple-talker lists with respect to the rate manipulation strongly suggests that talker variability affects some aspect of the rehearsal process. However, the comparisons made between the stimulus dimensions of talker variability and word confusability would not be as meaningful as they would be if the perceptual consequences of both dimensions were equivalent. To examine this, another experiment is now in progress in which we are collecting naming latencies for the stimulus words used in the present study. Our hopes are that the reaction time deficits caused by talker variability will approximate those caused by "hard" words.

In summary, the present data provide support for the explanation suggested by Martin, et al. that talker variability appears to affect recall of spoken word lists by influencing the speed and efficiency of the rehearsal processes used to transfer items to long-term memory. Casual observation and introspection suggests that humans perform most language related tasks with equal facility, despite superficial variations in voices or other sources of variability in the speech signal. As Martin et al. note, however, with more rigorous observation and experimentation, we find that our casual observations leave us blissfully ignorant of the complex processes underlying even our most "simple" behaviors.

---

[3] It is interesting that in another of Geiselman's studies (Geiselman, 1979), it was reported that subjects can *inhibit* the automatic encoding of voice information when such information interferes with a primary cognitive task. The findings of the present study, as well as the Martin et al. study do not seem to support this claim. However, it is possible that the deficits in recall for multiple-talker lists reported here could result from inefficient use of voice information, instead of from interfering effects of voice information. Geiselman's (1979) finding lends support to the notion that voice information is incorporated into the memory trace for the spoken word lists, not intentionally "stripped away" from the tokens during initial encoding.

References

Allard, F. & Henderson, L. (1976). Physical and name codes in auditory memory: The pursuit of an analogy. *Quarterly Journal of Experimental Psychology*, **28**, 475-482.

Cole, R.A., Coltheart, M., & Allard, F. (1974). Memory of a speaker's voice: Reaction time to same- or different-voiced letters. *Quarterly Journal of Experimental Psychology*, **26**, 1-7.

Craik, F.I.M., & Kirsner, K. (1974). The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology*, **26**, 274-284.

Fodor, J.A. (1983). *Modularity of mind.* Cambridge, MA: MIT Press.

Garner, W.R. (1974). *The processing of information and structure.* Potomac, MD: Erlbaum.

Geiselman, R.E. (1979). Inhibition of the automatic storage of speaker's voice. *Memory and Cognition*, **7**, 201-204.

Geiselman, R.E., & Bellezza, F.S. (1976). Long-term memory speaker's voice and source location. *Memory and Cognition*, **4**, 483-489.

Geiselman, R.E. & Crawley, J.M. (1983). Incidental processing of speaker characteristics: Voice as connotative information. *Journal of Verbal Learning and Verbal Behavior*, **22**, 15-23.

Geiselman, R.E., & Glenny, J. (1977). Effects of imagining speakers' voices on the retention of words presented visually. *Memory and Cognition*, **5**, 499-504.

Glanzer, M., & Cunitz, A.R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior*, **5**, 351-360.

House, A.S., Williams, C.E., Hecker, M.H.L., & Kryter, K.D. (1965). Articulation-testing methods: Consonantal differentiation with a closed-response set. *Journal of the Acoustical Society of America*, **37**, 158-166.

Kosslyn, S.M., & Matt, A.M.C. (1977). If you speak slowly, do people read your prose slowly? Person-particular speech recoding during reading. *Bulletin of the Psychonomic Society*, **9**, 250-252.

Kučera, F., & Francis, W. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Logan, J.S., & Pisoni, D.B. (1987). Talker variability and the recall of spoken word lists: A replication and extension. *Research on speech perception progress report no. 13.* Bloomington, IN: Indiana University.

Luce, P.A. (1986). Neighborhoods of words in the mental lexicon. *Research on speech perception technical report no. 6.* Bloomington, IN: Indiana University.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (in press). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Mullennix, J.W., & Pisoni, D.B. (1987). Talker variability effects and processing dependencies between word and voice. *Research on speech perception progress report no. 13.* Bloomington, IN: Indiana University.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (in press). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America.*

Murdock, B.B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, **64**, 482-488.

Nusbaum, H.C., Pisoni, D.B., & Davis, C.K. (1984). Sizing up the Hoosier mental lexicon: Measuring the familiarity of 20,000 words. *Research on speech perception progress report no. 10.* Bloomington, IN: Indiana University.

Peterson, L.J., & Peterson, M.J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, **58**, 193-198.

Rundus, D. (1971). Analysis of rehearsal processes in free recall. *Journal of Experimental Psychology*, **89**, 43-50.

Sumby, W.H. (1963). Word frequency and serial position effects. *Journal of Verbal Learning and Verbal Behavior*, **1**, 443-450.

Summerfield, Q., & Haggard, M.P. (1973). Vocal tract normalisation as demonstrated by reaction times. *Report on research in progress in speech perception*, **2**, Belfast, Northern Ireland: The Queen's University of Belfast.

III. PUBLICATIONS

Papers Published

Davis, S. (1988). Syllable onsets as a factor in stress rules. *Phonology*, *5*, 1-19.

Davis, S. (1988). Italian onset structure and the distribution of il and lo. *Proceedings of the Fourth Eastern States Conference on Linguistics* (pp. 64-74).

Davis, S. (1988). On the nature of internal reduplication. In *Theoretical Morphology* (pp. 305-323). Orlando, FL: Academic Press.

Gierut, J.A. (1988). Comparative research on language learning. *Language Learning*, **38**, 413-438.

Gierut, J.A. (1988). Review of *Language and Experience: Evidence From the Blind Child* by B. Landau and L.R. Gleitman. *Studies in Second Language Acquisition*, **10**, 115-118.

Gierut, J.A., & Pisoni, D.B. (1988). Speech perception. In N.Lass, L.McReynolds, J.Northern, and D.Yoder (Eds.), *Handbook of speech-language pathology and audiology* (pp. 264-287). Philadelphia: B.C. Decker.

Greene, B.G., & Pisoni, D.B. (1988). Perception of synthetic speech by adults and children. In L.E. Bernstein (Ed.), *The vocally impaired: Clinical practice and research* (pp. 206-248). Philadelphia, PA: Grune and Stratton.

Greenspan, S.L., Nusbaum, H.C., & Pisoni, D.B. (1988). Perceptual learning of synthetic speech produced by rule. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **14**, 421-433.

Pisoni, D.B. (1988). Review of *Speech and Speaker Recognition* by M.R. Schroeder. *Journal of the Acoustical Society of America*, **84**, 457-458.

Pisoni, D.B., Manous, L.M. & Dedina, M.J. (1987). Comprehension of natural and synthetic speech: Effects of predictability on the verification of sentences controlled for intelligibility. *Computer Speech and Language*, **2**, 303-320.

Stemberger, J.P. (1988). Between-word processes in child phonology. *Journal of Child Language*, **15**, 39-61.

Summers, W.V. (1988). F1 provides information for final-consonant voicing. *Journal of the Acoustical Society of America*, **84**, 485-492.

Summers, W.V., Pisoni, D.B., Bernacki, B., Pedlow, R.I., & Stokes, M.A. (1988). Effects of noise on speech production: Acoustic and perceptual analyses. *Journal of the Acoustical Society of America*, **84**, 917-928.

Manuscripts Accepted for Publication (In Press):

Charles-Luce, J. (in press). The effects of semantic context on voicing neutralization. *Phonetica.*

Charles-Luce, J., & Luce, P.A. (in press). Some structural characteristics of young children's lexicons. *Journal of Child Language.*

Charles-Luce, J., Luce, P.A., & Cluff, M. (in press). Retroactive influence of syllable neighborhoods. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Davis, S. (in press). Cross-vowel phonotactic constraints. *Computational Linguistics.*

Davis, S. (in press). On a non-argument for the rhyme. *Journal of Linguistics.*

Davis, S. (in press). The location of [continuant] in feature geometry. *Lingua.*

Davis, S., & Tsujimura, N. (in press). The morphophonemics of Japanese verbal conjugation: An autosegmental account. *Proceedings of the fifth eastern states conference on linguistics.*

Dedina, M.J., & Nusbaum, H.C. (in press). PRONOUNCE: A program for pronunciation of new words by analogy. *Computer Speech and Language.*

Gierut, J.A. (in press). Maximal opposition approach to phonological treatment. *Journal of Speech and Hearing Disorders.*

Gierut, J.A. (in press). Developing descriptions of phonological systems: A surrebuttal. *Applied Psycholinguistics.*

Gierut, J.A., & Pisoni, D.B. (in press). Speech perception. In J. Northern (Ed.), *Study guide for handbook of speech-language pathology and audiology.* Philadelphia: B.C. Decker.

Gierut, J.A., Elbert, M., & Dinnsen, D.A. (in press). Issues of linguistic analysis and experimental design: Reply to Diedrich. *Journal of Speech and Hearing Research.*

Goldinger, S.D., Luce, P.A., & Pisoni, D.B. (in press). Priming lexical neighbors of spoken words: Effects of competition and inhibition. *Journal of Memory and Language.*

Johnson, K. (in press). Higher formant normalization results from auditory integration of F2 and F3. *Perception and Psychophysics.*

Logan, J.S., Greene, B.G., & Pisoni, D.B. (in press). Segmental intelligibility of synthetic speech produced by rule. *Journal of the Acoustical Society of America.*

Luce, P. A. (in press). Similarity neighborhoods and word frequency effects in visual word identification: Sources of facilitation and inhibition. *Journal of Memory and Language.*

Luce, P.A., Pisoni, D.B., and Goldinger, S.D. (in press). Similarity neighborhoods of spoken words. In G. Altmann (Ed.), *Cognitive models of speech perception: Psycholinguistic and computational perspectives.* Cambridge, MA: MIT Press.

Martin, C.S., Mullennix, J.W., Pisoni, D.B., & Summers, W.V. (in press). Effects of talker variability on recall of spoken word lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition.*

Mullennix, J.W., & Pisoni, D.B. (in press). Speech perception: Analysis of biologically significant signals. In R.J. Dooling and S.H. Hulse (Eds.) *The comparative psychology of audition: Perceiving complex sounds.* Hillsdale, NJ: Erlbaum.

Mullennix, J.W., Pisoni, D.B., & Martin, C.S. (in press). Some effects of talker variability on spoken word recognition. *Journal of the Acoustical Society of America.*

Pisoni, D.B. (in press). Modes of processing speech and nonspeech signals. In I.G. Mattingly (Ed.), *Modularity and the motor theory of speech perception*. Hillsdale, NJ: Erlbaum.

Pisoni, D.B., & Martin, C.S. (in press). Effects of alcohol on the acoustic-phonetic properties of speech: Perceptual and acoustic analyses. *Alcoholism: Clinical and Experimental Research*.

Stemberger, J.P. (in press). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In B.Baars (Ed.), *The psychology of errors: A window on the mind?* New York, NY: Plenum.

IV. Speech Research Laboratory Staff, Faculty, and Technical Personnel

(1/1/88 - 12/31/88)

Research Personnel:

David B. Pisoni, Ph.D. ------------- Professor of Psychology and Director
Beth G. Greene, Ph.D. -------------- Research Scientist and Assoc. Director *
Jan Charles-Luce, Ph.D. ------------ Research Associate †
Paul A. Luce, Ph.D. ---------------- Research Associate ‡
James V. Ralston, Ph.D. ------------ Visiting Asst. Professor of Psychology
W. Van Summers, Ph.D. -------------- Research Associate

Stuart A. Davis, Ph.D. ------------- NIH Post-doctoral Trainee
Keith A. Johnson, Ph.D. ------------ NIH Post-doctoral Trainee
John W. Mullennix, Ph.D. ----------- NIH Post-doctoral Trainee

Michael S. Cluff, B.S. ------------- NIH Pre-doctoral Trainee

Stephen D. Goldinger, B.A. --------- Graduate Research Assistant
Mary Jo Lewellen, M.A.-------------- Graduate Research Assistant
Nancy L. Lightfoot, B.A. ----------- Graduate Research Assistant
John S. Logan, B.S. ---------------- Graduate Research Assistant
Heng Jie Ma, B.A. ------------------ Graduate Research Assistant
Joanne K. Marcario, B.S. ----------- Graduate Research Assistant
Christopher S. Martin, B.A. -------- Graduate Research Assistant
Robert I. Pedlow, M.Sc. ------------ Graduate Research Assistant
Michael A. Stokes, B.A. ------------ Graduate Research Assistant

---

* Also, Center for Reading and Language Studies, School of Education.

† Now at Dept. of Communicative Disorders and Sciences, SUNY at Buffalo, Buffalo, NY.

‡ Now at Dept. of Psychology, SUNY at Buffalo, Buffalo, NY.

Technical Support Personnel:

Robert H. Bernacki, B.A. ----------- Research Analyst
Cheryl L. Blackerby ---------------- Administrative Secretary
Michael J. Dedina, M.S. ------------ Research Programmer/Assistant *
Dennis M. Feaster, B.A. ------------ Software Development Specialist
Jerry C. Forshee, M.A. ------------- Computer Systems Analyst
Luis R. Hernandez, B.A. ------------ Software Development Specialist
David A. Link ---------------------- Electronics Engineer
Gary Link -------------------------- Technical Assistant

Denise Beike ----------------------- Undergraduate Research Assistant
Lisa Jalbert ----------------------- Undergraduate Research Assistant
Amy Lawlor ------------------------- Undergraduate Research Assistant
Scott Lively ----------------------- Undergraduate Research Assistant
Bridget Robinson ------------------- Undergraduate Research Assistant

---

* Now at IBM Corporation, San Jose, CA.